Enhancement of LC-MS Proteomic Analysis through Novel Data Acquisition Strategies

by Simion Kreimer

B.S. in Cognitive Psychology, University of California, Davis
M.S. in Biopharmaceutical Regulatory Sciences, Northeastern University

A dissertation submitted to

The Faculty of
The College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

December 6, 2016

Dissertation directed by

Jeff Agar
Associate Professor of Chemistry and Chemical Biology
Acknowledgements

I would like to sincerely thank all of the excellent colleagues and friends who made this dissertation possible. First, and foremost, I would like to thank my advisor, Professor Barry Karger, for his guidance, support, and patience. I began learning from Professor Karger in his graduate course on separations and continued as a research assistant in his group. His decades of experience were an inspiration and invaluable resource in my development. For my thesis defense, Professor Karger could not be my official advisor due to his Emeritus status as of July 1, 2016, and I am grateful to Professor Jeff Agar for stepping into the role. Next, I would like to thank Dr. Alexander Ivanov for his training and assistance in the research presented in Chapters 2 and 4. Dr. Ivanov’s adventurous attitude in exploring novel research topics and technologies is a trait that I hope to carry on in my future scientific endeavors.

I would also like to express my gratitude to the members of the Karger group for their friendship and contributions. Enormous thanks to Dr. Yuanwei “Abby” Gao who contributed significantly to the development of the host-cell protein analysis workflow and Somak Ray who helped with automation of the data analysis presented in Chapter 3. Special thanks to Dr. David Bush and Dr. Shujia “Daniel” Dai who, as post-docs in the group, were always generous in sharing their experience and knowledge. Further thanks to the other members of the Karger group: Arseniy Belov, Dr. Siyuan Liu, Dr. Xianzhe Wang, Dr. Krishan Kumar, Dr. Vennela Mullangi, Dr. Wenqin Ni, Dr. Chen Li, Dr. Zhenke Liu, and Dr. Siyang Li.

I would like to thank the collaborators outside of the Barnett Institute for their contributions to the presented research. The intelligent data acquisition platform presented in Chapter 2 was co-developed with Dr. Mikhail Belov and Bill Danielson from Spectroglyph (Kennewick, WA). The database search integrated into the data acquisition platform was developed by Lev Levitsky from the Institute of Energy Problems of Chemical Physics (Moscow, Russia). The workflow for analysis of host-cell proteins presented in Chapter 3 was developed in collaboration with Dr. Mi Jin (BMS, now at Teva), Dr. Nesredin Mussa, Dr. Zhijun Tan, Dr. Li Tao, and Dr. Zhengjian Li from Bristol-Myers Squibb (Devens, MA) who provided the samples. The investigation of extracellular vesicles presented in Chapter 4 was carried out with help from Dr. Ionita Ghiran’s group and Dr. John Tigges at Beth Israel Medical Center (Boston, MA), and Dr. Natasha Barteneva at Harvard Medical School (Boston, MA) who provided invaluable expertise,
samples, and access to instrumentation. Further thanks to Dr. David Frank and Dr. Jennifer Yeh at the Dana Farber Cancer Institute for our collaboration to identify novel STAT3 interacting proteins.

I would also like to thank Professor Roger Giese and Professor Zhaohui "Sunny" Zhou for their participation on my dissertation committee. Finally, I would like to thank Professor Ira Krull who convinced me to pursue a Doctorate when I was working on a Master's degree, and continued to support me throughout the journey.
Dedication

This dissertation is dedicated to my family, immediate and extended.
Abstract of Dissertation

This dissertation is focused on optimization of data acquisition and processing in bottom-up LC-MS protein analysis. The first chapter introduces LC-MS proteomics and describes the three fundamental data acquisition strategies: targeted MS, data-independent acquisition (DIA), and data-dependent acquisition (DDA). The Top N DDA scheme in which the N highest intensity eligible precursors are individually fragmented during each data acquisition cycle is widely implemented despite high intensity ion bias, stochastic identifications, and redundant precursor sampling between experiments leading to a low plateau in identifications, known as saturation. In Chapter 2, more sophisticated DDA precursor ion selection algorithms were developed to reduce redundant precursor sampling. The applied algorithms significantly increased identification of low abundance peptides and proteins beyond saturation of the conventional approach in a proof-of-concept analysis of HeLa lysate. In Chapter 3, DIA in which all precursors are repeatedly fragmented in parallel throughout elution and targeted MS in which a set of peptides is monitored individually throughout elution were combined in host cell protein characterization (HCP). While DDA poorly identifies low intensity proteins in a high background, 24 HCPs were reproducibly identified and quantitated down to low part per million levels in a purified monoclonal antibody using the developed workflow. In Chapter 4, the reproducibility of DIA was used to improve on the stochastic protein identifications by DDA during comparison of heterogeneous extracellular vesicle samples isolated from plasma by size exclusion chromatography. Summarily, this dissertation illustrates potential for continual improvement in LC-MS analytical capabilities through intelligent instrument control and data processing.
## Table of Contents

Acknowledgements .............................................................................................................. ii

Dedication ............................................................................................................................... iv

Abstract of Dissertation......................................................................................................... v

List of Figures.......................................................................................................................... x

List of Tables............................................................................................................................ xii

List of Abbreviations ............................................................................................................. xiii

Chapter 1: Introduction to LC-MS Data Analysis and Proteomics ........................................... 1

1.1 Abstract ............................................................................................................................. 2

1.2 Overview of mass spectrometry ...................................................................................... 2

1.3 Overview of liquid chromatography ................................................................................. 3

1.4 Overview of data acquisition strategies in LC-MS ............................................................ 8

1.4.1 Targeted mass spectrometry ....................................................................................... 8

1.4.2 Data dependent acquisition (DDA) ............................................................................ 9

1.4.3 Data independent acquisition (DIA) .......................................................................... 10

1.5 Mass spectrometry instrumentation ................................................................................. 13

1.5.1 The Orbitrap mass analyzer ....................................................................................... 13

1.5.2 LTQ Orbitrap XL ......................................................................................................... 15

1.5.3 QExactive and QExactive Plus ................................................................................... 16

1.6 LC-MS based protein analysis ........................................................................................ 17

1.6.1 Proteins as substrates for LC-MS analysis .................................................................. 17

1.6.2 Applications of LC-MS based proteomic analysis ....................................................... 18

1.6.3 Bottom-up proteomics ............................................................................................... 19

1.6.4 Reversed phase LC in bottom-up proteomics ............................................................... 20

1.7 Bottom-up proteomic data analysis ................................................................................. 24

1.7.1 DDA data analysis .................................................................................................... 25

1.7.2 Evaluation of search result confidence by target-decoy ............................................. 27

1.7.3 DIA data analysis ..................................................................................................... 28

1.7.4 Complications in bottom-up data analysis ................................................................. 30
# Chapter 1: Quantitative LC-MS Proteomics

1.8 Quantitative LC-MS proteomics .......................................................... 31

1.8.1 Stable isotope labeled quantitation .................................................. 32

1.8.2 High throughput label-free quantitation ......................................... 36

1.8.3 Targeted label-free quantitation ..................................................... 36

1.8.4 Accurate and absolute quantitation by isotope dilution MS ............... 37

1.9 Conclusion ......................................................................................... 38

1.10 References ....................................................................................... 40

## Chapter 2: Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling

2.1 Abstract............................................................................................... 49

2.2 Introduction ........................................................................................ 49

2.3 Materials and methods ....................................................................... 52

2.3.1 Materials ....................................................................................... 52

2.3.2 Liquid chromatography ................................................................. 53

2.3.3 Mass spectrometry ......................................................................... 53

2.3.4 Exclusion of identified peptides ..................................................... 54

2.3.5 Data analysis .................................................................................. 54

2.4 Results and Discussion ....................................................................... 55

2.4.1 Precursor selection algorithms applied through Smart MS$^2$ ............ 55

2.4.2 Performance metrics of Smart MS$^2$ and conventional DDA .......... 59

2.4.3 Effects of iterative exclusion (Strategy 2) on precursor selection ........ 61

2.4.4 Differences in identifications between Strategies 1 and 2 ................. 66

2.4.5 Iterative exclusion of identified precursors vs. all fragmented precursors 68

2.5 Conclusion .......................................................................................... 69

2.6 Assessment of Xcalibur driven exclusion ........................................... 71

2.7 Acknowledgements ............................................................................. 72

2.8 Reference ............................................................................................ 73

## Chapter 3: Comprehensive HCP Profiling by Targeted and Untargeted Analysis of DIA Mass Spectrometry Data with PRM Verification

3.1 Abstract............................................................................................... 77
3.2 Introduction ...........................................................................................................77
3.3 Materials and Methods ........................................................................................80
3.3.1 Materials and equipment ..............................................................................80
3.3.2 Sample preparation .........................................................................................81
3.3.3 LC-MS .............................................................................................................81
3.3.4 Targeted assay library assembly ....................................................................81
3.3.5 DIA to PRM analysis ......................................................................................83
3.4 Results and Discussion .......................................................................................84
3.4.1 Overview of the DIA to PRM workflow for HCP identification and quantitation ...84
3.4.2 Stage 1: targeted assay library assembly .......................................................85
3.4.3 Stages 2 and 3: DIA analysis of the purified mAb .........................................87
3.4.4 Stage 4A: PRM verification ............................................................................89
3.4.5 Stage 4B: PRM quantiation ..........................................................................92
3.5 Conclusion ...........................................................................................................94
3.6 Detailed description of targeted assay library assembly and DIA data processing ...95
3.6.1 Targeted assay library assembler ..................................................................96
3.6.1.1 Required software .....................................................................................96
3.6.1.2 Script input ................................................................................................96
3.6.1.3 Script output ...............................................................................................96
3.6.1.4 Script procedure .......................................................................................98
3.6.2 DIA data processing .....................................................................................99
3.6.2.1 Required software .....................................................................................99
3.6.2.2 Script input ................................................................................................99
3.6.2.3 Script output ...............................................................................................100
3.6.2.4 Script procedure .......................................................................................102
3.7 References ..........................................................................................................105

Chapter 4: Combination of DDA and DIA in a Preliminary Investigation of Extracellular Vesicles Extracted from Blood Plasma by Size Exclusion Chromatography .........................................................109
4.1 Abstract .............................................................................................................110
## Table of Contents

4.2 Introduction ............................................................................................................................................. 111

4.2.1 Nomenclature and EV categories ........................................................................................................... 111

4.2.2 Isolation of EVs and potential sources of contamination .............................................................................. 112

4.2.3 Proteomic profiling of EVs .......................................................................................................................... 118

4.2.4 SEC based EV isolation ............................................................................................................................... 119

4.3 Materials and Methods ................................................................................................................................. 120

4.3.1 Materials .................................................................................................................................................. 120

4.3.2 Methods .................................................................................................................................................... 121

4.3.2.1 Preparation of single-use SEC columns ................................................................................................. 121

4.3.2.2 SEC isolation of EVs ............................................................................................................................... 121

4.3.2.3 Nano-flow cytometric analysis of SEC fractions ..................................................................................... 122

4.3.2.4 Processing of EV containing SEC fractions ......................................................................................... 122

4.3.2.5 Nano-LC-MS analysis ............................................................................................................................ 123

4.3.2.6 Peptide and protein identification ......................................................................................................... 124

4.3.2.7 Experimental design ................................................................................................................................ 125

4.4 Results and Discussion ................................................................................................................................. 125

4.4.1 SEC enriches EVs and depletes serum proteins ....................................................................................... 125

4.4.2 DDA proteomic analysis of SEC isolated EVs .......................................................................................... 127

4.4.3 Complementary DIA analysis .................................................................................................................... 130

4.4.4 Future directions ....................................................................................................................................... 133

4.5 Conclusion .................................................................................................................................................. 134

4.6 References .................................................................................................................................................. 135
List of Figures

Figure 1.1 – Three dimensional rendering of LC-MS data............................................................................. 7
Figure 1.2 – Selection of precursors: DDA vs. DIA. ...................................................................................... 11
Figure 1.3 – C-trap and Orbitrap illustration. .................................................................................................. 14
Figure 1.4 – LTQ Orbitrap XL Schematic. ....................................................................................................... 15
Figure 1.5 – QExactive Plus schematic. .......................................................................................................... 17
Figure 1.6 – Increased peak capacity with UPLC. ............................................................................................ 22
Figure 1.7 – High - low pH reversed phase LC separation.................................................................................. 24
Figure 1.8 – Illustration of DDA and DIA data. ................................................................................................. 25
Figure 1.9 – Typical score distribution of decoy and target peptide spectral matches...................................... 28
Figure 1.10 – 6-plex TMT reagents and workflow. .......................................................................................... 35
Figure 2.1 – Smart MS\(^2\) flow control diagram. ........................................................................................... 56
Figure 2.2 – PCI filter (A) and on-the-fly exclusion list alignment (B). ................................................................. 59
Figure 2.3 – Effects of iterative exclusion of precursor selection....................................................................... 65
Figure 2.4 – Comparison of identifications: Strategy 1 vs. Strategy 2................................................................. 66
Figure 2.5 – Comparison of identifications: Strategy 3 vs. Strategy 2................................................................. 68
Figure 2.6 – Xcalibur driven exclusion does not improve peptide identifications............................................. 72
Figure 3.1 – Workflow for HCP identification and quantitation. ....................................................................... 85
Figure 3.2 – Comparison of protein intensity and number of assays in the targeted assay library. ................. 87
Figure 3.3 – Putative and verified peptide identifications from 3 DIA runs. ..................................................... 89
Figure 3.4 – Flow diagram for automatic assembly of targeted assay library.................................................. 97
Figure 3.5 - Flow diagram for automated DIA data processing. ...................................................................... 101
List of Tables

Table 1.1 – Comparison of data acquisition strategies................................................................. 12

Table 1.2 – Comparison of DIA strategies.................................................................................. 12

Table 1.3 – Comparison of quantitative proteomic strategies.................................................. 32

Table 2.1 – Precursor selection logic implemented in each data acquisition strategy............. 60

Table 2.2 – Performance metrics for each data acquisition strategy. ...................................... 63

Table 2.3 – Overlap in peptide identifications between replicate LC-MS experiments. .......... 64

Table 3.2 – Testing of PRM quantiation with protein standards .............................................. 94

Table 4.1 – Potential Non-EV contaminants.............................................................................. 117

Table 4.2 – Fold changes and p-scores of potentially differentiated proteins.......................... 130
List of Abbreviations

DDA – data dependent acquisition

DIA – data independent acquisition

LC – liquid chromatography

m/z – mass to charge

MS – mass spectrometry

MS1 – mass spectrometry precursor scan

MS2 – mass spectrometry fragment scan

MSn – mass spectrometry scan after (n – 1) rounds of fragmentation

PRM – parallel reaction monitoring

PTM – post-translation modification

SEC – size exclusion chromatography

HILIC – hydrophilic liquid interaction chromatography

RP – reversed phase

CID – collision induced dissociation

HCD – high energy collision induced dissociation

FWHM – full width half at maximum height/intensity

TP – true positive

TN – true negative

FP – false positive

FN – false negative
SIL – stable isotope labeled

AGC – automatic gain control

LTQ - Linear ion trap quadrupole

FT - Fourier transform

eFT - enhanced Fourier Transform
Chapter 1: Introduction to LC-MS Data Analysis and Proteomics
1.1 Abstract

This introductory chapter defines the data generated in liquid chromatography paired to tandem mass-spectrometry (LC-MS) experiments. Initially, LC-MS data and data acquisition strategies are described in general terms, and the schematics of the MS instruments used in this dissertation presented. The discussion then moves to the implementation of the instrumentation and data acquisition in LC-MS based proteomic analysis. Next, the strategies for automatic processing of LC-MS data to identify and quantitate proteins are introduced. The chapter is concluded with an overview of the subsequent chapters and the publications resulting from the presented investigations.

1.2 Overview of mass-spectrometry

Mass spectrometry (MS) is an analytical technique in which the mass to charge ratio (m/z) of ions in the gaseous phase is measured. The intensities of m/z signals are proportional to the number of the corresponding analyte ions injected into the mass analyzer. Due to naturally occurring elemental isotopes, analytes are observed as clusters of m/z values separated by the mass of a neutron (1 Dalton) divided by the analyte’s charge, with the relative intensities of isotopic peaks proportional to the natural abundance of the isotopes. MS resolution is measured as the m/z value divided by the width of the signal peak at half of the maximum intensity (FWHM). Modern mass spectrometers have sufficient resolution to detect adjacent isotopic peaks and consequently determine the charge and ionized mass of the analytes. This information, with sufficient mass accuracy and resolution, can be used to compute the elemental composition of smaller analytes. For example, mass measurements with the error below 5 part per million (ppm) can be used to unambiguously determine the elemental composition of analytes below 200 m/z. However, intact mass measurements alone cannot distinguish between isobaric species (identical mass but different arrangements of atoms).

The structure of an analyte can be characterized through fragmentation and detection of the product ions of the analyte in the mass spectrometer. Measurement of the intact analytes’ m/z is referred to as the precursor or MS1 scan, measurement of the fragment ions’ m/z is referred to as the product, or MS2 scan, and analysis where both MS1 and MS2 scans are acquired is commonly referred to as MS/MS or
tandem MS. Structural details can be deduced from the fragment masses and prior knowledge of precursor fragmentation pathways. Some analytes such as peptides can be, to a large extent, unambiguously identified from MS1 and MS2 measurements, while other analytes may require multiple rounds of fragmentation (MSn) and/or orthogonal analysis (e.g. infrared spectroscopy). It is important to note that while tandem MS can determine the linkage of components, conventional MS cannot determine chirality or the spatial arrangement of atoms. Although, ion mobility MS in which low energy collision with an inert gas is used to separate ions by collisional cross-section can, in some instances, disambiguate stereoisomers (two compounds which differ only in the spatial arrangements of atoms).

MS scans are generally acquired on a millisecond time scale, enabling unparalleled throughput in identification of individual components within a mixture. The MS1 scan simultaneously measures the precursor m/z of all analytes within the relevant m/z range, and an MS2 scan can be acquired for each detected analyte to provide additional details for identification. However, in mixtures of a vast number of components, commonly referred to as complex mixtures, the MS1 spectrum is dominated by the highest concentration and most efficiently ionized species while low abundance or poorly ionized species are masked. Masking occurs because of ion suppression, where the dominant analytes acquire the majority of the charge in the electrospray ionization process, and the ion collection mechanism of the mass analyzer becomes overwhelmed by the dominant ions. Additionally, the mass analyzer resolution may be insufficient to distinguish between ions of nearly identical m/z, resulting in spectral crowding which further reduces the ability to identify the lower intensity species. Separation of complex mixtures into more manageable sub-populations addresses these issues, greatly improves depth of analysis, and is achieved by pairing chromatographic or electrophoretic separation to mass spectrometry.

1.3 Overview of liquid chromatography

Liquid chromatography (LC) is the spatial separation of analytes in the liquid phase based on physical properties. An LC column contains a stationary solid phase and a mobile liquid phase. The analytes, which are dissolved in the mobile phase, are separated based on interaction with the stationary phase as the mobile phase flows through the column. Affinity for the stationary phase causes analytes to elute in
the order dictated by the strength of that interaction. The solute bands diffuse along the length of the column during the separation resulting in a roughly Gaussian (actually Poisson) distribution, commonly referred to as a chromatographic peak. Various LC modes exploit different physical or chemical properties: size exclusion chromatography (SEC) separates analytes by size, ion exchange chromatography by charge at a given pH, hydrophilic interaction liquid chromatography (HILIC) by hydrogen bonding, and reversed phase (RP) chromatography by “hydrophobicity”, or binding to a hydrophobic solid phase. There are also mix-mode chromatography techniques which combine multiple retention factors in the separation mechanism. Furthermore, affinity chromatography uses specific or selective interactions between the analyte and the substrate to retain analytes based on availability of a binding motif.

SEC, also referred to as gel-filtration chromatography, is implemented in Chapter 4 and is used here to illustrate basic LC concepts. In SEC, a low-binding stationary phase with a heterogeneous pore distribution is used to separate analytes by size, or more accurately hydrodynamic volume. Polar, ionic, and hydrophobic interactions between the analytes and the stationary phase are minimized. Analytes are only retained through pore occupancy: outside of the pores the analytes are carried along with the mobile phase, but are stagnant relative to the mobile phase when they diffuse into the pores. Small analytes are able to diffuse into large and small pores, while larger analytes are only able to occupy the large pores and are thus carried through the column quicker. SEC is relatively low resolution (defined below) and is not typically paired with MS.

RP and HILIC are two LC techniques which are readily coupled on-line to mass spectrometry (LC-MS) through electrospray ionization (ESI). In ESI, the analytes eluting off the analytical column are nebulized under an electrostatic potential. The solvent evaporates from the sprayed droplets, with or without the assistance of an inert sheath gas. As the droplets shrink the concentration of ions increases, eventually resulting in fission when the Coulombic repulsion from the concentrated charge exceeds the droplet surface tension; this threshold is referred to as the Rayleigh limit. Sequential fission produces progressively finer droplets until the ionized analytes are ejected into the gas phase. The droplet surface tension and consequently Rayleigh limit are dependent on the solvent composition and the efficiency of
ESI (ratio of analytes ions in the gas phase compared to analytes eluting from the column) is higher in solvents which readily evaporate. ESI efficiency is also dependent on the LC flow rate because slower flow rates generate smaller droplets that ionize readily and reduce matrix effects. Thus high sensitivity analysis is achieved with the use of narrow bore LC columns (internal diameter below 300 micrometers) at nano-liters per minute flow rates.\(^{20}\)

Complex samples can be separated into less complex populations by LC, which are nebulized and ionized as they elute off the column, and analyzed by MS. This ability to physically isolate low abundance components from high abundance species greatly increases the number of solutes that can be identified by MS, because losses of sensitivity due to ion suppression and spectral crowding are reduced.\(^{21}\) Thus, the key consideration in pairing LC to MS is the maximum number of separate populations (observed as peaks), or peak capacity, generated during the separation.\(^{22}\) Peak capacity is defined as the duration of the separation divided by the average peak width,\(^{23}\) and this metric is closely related to LC resolution.\(^{24}\)

LC resolution (\(R_s\)) is defined as the difference of adjacent analyte retention times (\(T_{R1}\) and \(T_{R2}\)) divided by the average of the two peak baseline widths (\(w_1\) and \(w_2\)).\(^{25}\)

\[
R_s = \frac{T_{R1} - T_{R2}}{\frac{1}{2}(w_1 + w_2)} \quad \text{(Equation 1)}
\]

LC resolution is dependent on the square root of column efficiency also referred to as the number of theoretical plates (\(N\)), selectivity (\(\alpha\)), and retention factor (\(k\)).\(^{25}\)

\[
R_s = \frac{\sqrt{N}}{4} \left( \frac{a-1}{\alpha} \right) \left( \frac{k}{k+1} \right) \quad \text{(Equation 2)}
\]

The selectivity and retention factor are based on the interactions of the analyte with the stationary phase under the given mobile phase composition and temperature. The column efficiency, \(N\), is dependent on the structural properties of the solute and column and the velocity of the mobile phase (\(u\)). The number of theoretical plates (\(N\)) is equal to the length of the column (\(L\)) divided by the theoretical plate height (HETP). HETP is a measure of band broadening, low HETP means narrow peaks, which translates to high resolution, high peak capacity, and improved LC-MS performance. In LC columns where the
stationary phase is composed of tightly packed microscopic beads, HETP is modeled by the Van Deemter equation:26

\[ HETP = \frac{L}{N} = a(d_p) + \frac{b}{u} + c(d_p)^2 u \] (Equation 3)

The first term, \( a(d_p) \), which represents band broadening due to non-uniformity of flow through and around the beads, is proportional to the particle diameter \( (d_p) \). Smaller particles (low \( d_p \)) are packed more uniformly resulting in more uniform flow through the column. The second term, \( b/u \), represents longitudinal diffusion of the analyte and is inversely proportional to the flow velocity \( (u) \). The constant \( b \) represents the diffusion of the analyte in the mobile phase. The third term, \( c(d_p)^2 u \), is proportional to the particle diameter and the square of flow velocity and represents band broadening due to diffusion of the analyte population within the stagnant mobile phase in the particle pores. The analyte band broadens when a portion of the population diffuses into the particle pores and becomes temporarily stationary relative to the remainder of the population which is carried forward by the mobile phase flow. In smaller particles with shallower pores, the analytes spend less time in the pores thus reducing band broadening due to this effect.27 The Van Deemter equation is used to identify the optimal flow rate, which produces the lowest theoretical plate height and highest resolution based on experimentally determined \( a, b, \) and \( c \) constants. High resolution translates to narrow peaks and separation of analytes into more sub-populations, enabling MS to identify more species. Additionally, LC retention time can be used as an identifying characteristic of analytes.26 The LC techniques used in this dissertation will be further discussed in section 1.5.

Typical high resolution LC peaks are on the ~10 second time scale, while MS scans are acquired on the ~0.1 second time scale, which means that an MS instrument is able to trace chromatographic peaks using multiple measurements. Figure 1.1 presents data generated by repeated MS1 scans acquired over the course of an LC separation rendered in three dimensions using the OpenMS TOPP viewer,28,29 with LC retention time on the X-axis, the m/z on the Z-axis, and ion intensity on the Y-axis. Figure 1.1 presents a minor segment of a typical LC-MS run to emphasize the Gaussian shape of the chromatographic peaks. In the complex samples discussed in this dissertation more than 100,000 components are typically detected over 100 minutes of separation requiring at least 25 MS2
scans/second to fragment each individual analyte. This is not yet feasible due to limited MS analyzer sampling speed (scans/second). Furthermore, the quality of an MS2 spectrum correlates with the number of analyte ions that are fragmented and low intensity precursors increase the accumulation times required to generate an informative MS2 spectrum, further reducing scan speed. This dissertation is focused on optimization of MS2 spectra acquisition to best address the experimental requirements given the limited instrument scan speed and the currently available data acquisition strategies are introduced in the following section.

**Figure 1.1 – Three dimensional rendering of LC-MS data.**
MS1 trace of several precursors in the 645-658 m/z range over several minutes.
1.4 Overview of data acquisition strategies in LC-MS

There are three fundamental approaches to selection of precursors for MS2 in LC-MS: (1) Specification of m/z range(s) for fragmentation prior to the experiment, also referred to targeted analysis, which includes selected reaction monitoring, multiple reaction monitoring, and parallel reaction monitoring (PRM). (2) On-the-fly selection and fragmentation of individual precursor ions identified in the MS1 scan based on a set of rules, also referred to as data dependent acquisition (DDA). (3) Selection and fragmentation of all ions within the relevant m/z range independently of intensity through systematic scanning with wide isolation windows and co-fragmentation of multiple precursors, also known as data independent acquisition (DIA). A comparison of these data acquisition strategies is presented in Table 1.1 at the end of the section.

1.4.1 Targeted mass spectrometry

In targeted MS, analytes of interest are manually selected, and the mass spectrometer is programmed to fragment the corresponding m/z windows either for the entire duration of the experiment or for set retention time intervals. Targeted analysis is quantitative because the chromatographic peak can be traced from the multiple data-points acquired throughout analyte elution, and the integrated peak area is proportional to the concentration of the analyte in the sample. Quantitation is possible from integrating the peak of the intact analyte (MS1 peak area), but quantitation based on fragment area is more specific, because the transition of the intact precursor ion into a set of characteristic MS2 fragments distinguishes between isobaric species. The ability to accurately trace the chromatographic peak is dependent on the number of acquired data-points across the peak, and this limits the number of analytes which can be monitored during an experiment.

Throughput (number of analytes quantitated per experiment) is substantially increased when m/z ranges are only monitored during the corresponding analytes’ elution, but the retention times of these analytes must be known to perform the analysis. The retention time window over which an m/z range is repeatedly fragmented is dependent on the reproducibility of the LC retention time and peak width. Narrowing the retention time windows allows the mass spectrometer to monitor more analytes over an
LC-MS run, but the window must capture the entire elution of the analyte for accurate quantitation. Targeted LC-MS throughput is further increased using “triggered” targeted acquisition, in which the instrument intelligently waits for the start of analyte elution to acquire data for the corresponding m/z window. Nonetheless, targeted LC-MS is intended for accurate and reproducible quantitative analysis of a small set of precursors and is low-throughput when compared to DDA and DIA analysis. Quantitative analysis through targeted MS is discussed further in Section 1.8.

1.4.2 Data dependent acquisition (DDA)

The goal of DDA is to acquire high quality MS2 spectra for as many precursors as possible. Simple logical rules are used for automatic interpretation of the MS1 spectrum and selection of precursors to fragment. First, analytes are interpreted from the MS1 spectrum. Rules governing the interpretation of m/z intensities as analytes (instead of noise) include a set intensity threshold, detection of multiple isotopic peaks, and charge state criteria (e.g. $+1 < z < +6$). In modern instruments, the spectra are also de-convolved, meaning that m/z values that can be attributed to a single analyte, including different isotopic and charge states, are reduced to a single entity. In the most commonly used DDA strategy, referred to as Top N, the N number of highest intensity analyte precursor ions are selected for fragmentation from each MS1 scan. To avoid redundant fragmentation of the same high intensity precursors after each MS1 scan, an algorithm called dynamic exclusion is used to prohibit the selection of a recently fragmented precursor for a set time interval. With dynamic exclusion, the mass spectrometer is free to sample other precursors instead of focusing on a minority of high intensity precursors. Top N is a sensible strategy because higher intensity precursors are more likely to contribute sufficient ion current to produce an informative MS2 spectrum. However, Top N tends to focus on high intensity precursors leaving lower intensity species unidentified. A more in-depth discussion of DDA and potential improvements is presented in Chapter 2.
1.4.3 Data independent acquisition (DIA)

Data independent acquisition (DIA) is becoming an increasingly popular alternative to DDA. The key advantage of DIA is that fragmentation data are acquired for all precursors without the high intensity bias of Top N DDA. In DIA, MS2 spectra are acquired systematically and independently of the precursor scan by simultaneous fragmentation of multiple precursors using relatively large isolation windows (e.g. 25 m/z compared to 2 m/z in DDA). The differences between DDA and DIA is further clarified in Figure 1.2. There are four major variants of DIA: MS-E, SWATH, MSX-DIA, and pSMART, and their differences are summarized in Table 1.2. These variants are further customized by selection of isolation window size, scan resolution, and other parameters, but the key consideration in DIA is the compromise between isolation window specificity, the m/z range of analysis, and the average number of fragmentation spectra acquired over the elution of an analyte. Narrower isolation windows are more specific because they co-fragment fewer analytes, but they also cover a smaller portion of the m/z range, result in longer cycle times (the time required to scan the m/z range), and fewer data-points acquired across the analyte elution profile. DIA MS2 spectra are more crowded and complex than DDA MS2 spectra because each MS2 scan contains fragments from multiple precursors, but due to periodic fragmentation of each analyte, the chromatographic elution of fragment ions can be traced along with the precursor ions. Thus DIA data processing is more complicated and is further discussed in Section 1.6 and Chapter 3.
Figure 1.2 – Selection of precursors: DDA vs. DIA.
This simplified comparison of precursor ion fragmentation in DDA (top) and DIA (bottom) shows that in DDA the instrument selects the highest intensity precursors and fragments them individually and in DIA the instrument selects wide m/z ranges and fragments all precursors independent of intensity.
<table>
<thead>
<tr>
<th></th>
<th>DIA</th>
<th>DDA</th>
<th>PRM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Coverage</strong></td>
<td>All precursors are fragmented.</td>
<td>A stochastic subset of precursors is targeted for fragmentation.</td>
<td>Only a pre-selected set of precursors is fragmented.</td>
</tr>
<tr>
<td><strong>Complexity of Spectra</strong></td>
<td>Wide isolation windows select multiple precursors increasing complexity.</td>
<td>Narrow isolation windows select a single precursor.</td>
<td>Narrow isolation windows repeatedly select a single precursor.</td>
</tr>
<tr>
<td><strong>Analyte Identification</strong></td>
<td>Coincidence of precursor and characteristic fragment elution.</td>
<td>Intact mass and fragments in the generated spectrum.</td>
<td>Co-elution of characteristic fragments confirms precursor.</td>
</tr>
<tr>
<td><strong>Quantitation</strong></td>
<td>Precursor (MS1) peak area. Fragment (MS2) peak area.</td>
<td>Precursor (MS1) peak area. Spectral counts.</td>
<td>Precursor (MS1) peak area. Fragment (MS2) peak area.</td>
</tr>
<tr>
<td></td>
<td>SIL standards.</td>
<td>Isobaric mass tags.</td>
<td>SIL standards.</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>Sensitivity is reduced due to precursor co-isolation but all precursors are fragmented.</td>
<td>Sensitive for precursors which meet selection criteria, only fragmented precursors identified.</td>
<td>Maximum sensitivity for targeted precursors.</td>
</tr>
</tbody>
</table>

Table 1.1 – Comparison of data acquisition strategies.

<table>
<thead>
<tr>
<th>MS&lt;sup&gt;E&lt;/sup&gt;</th>
<th>SWATH</th>
<th>pSMART / vSWATH</th>
<th>MSX-DIA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>MS1 scan alternating with MS2 scan of all precursors co-fragmented.</td>
<td>SWATH with variable width windows based on precursor density.</td>
<td>Multiplexed fragmentation of randomized narrow isolation windows</td>
</tr>
<tr>
<td><strong>Isolation windows</strong></td>
<td>Entire m/z range of interest (largest)</td>
<td>Wide windows in sparse m/z ranges, narrow in dense.</td>
<td>multiple non-sequential (~5 m/z) windows co-fragmented</td>
</tr>
<tr>
<td><strong>Cycle time</strong></td>
<td>Fastest</td>
<td>Depends on isolation window widths.</td>
<td>Randomization of windows results in slowest cycle.</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>Maximum number of data points acquired for precursors and fragments.</td>
<td>Specificity is increased in high density m/z ranges compared to MS&lt;sup&gt;E&lt;/sup&gt;.</td>
<td>Highest specificity.</td>
</tr>
<tr>
<td><strong>Limitations</strong></td>
<td>Most complex spectra. Lowest sensitivity.</td>
<td>Can be improved further using variable isolation windows.</td>
<td>Requires customized instrument control for some vendors.</td>
</tr>
</tbody>
</table>

Table 1.2 – Comparison of DIA strategies.
Above are the strategies by which precursors ions can be selected for fragmentation in MS instruments. The following section explains how these strategies were implemented in the investigations presented in this dissertation by providing schematics of the instrumentation and a description of the ion manipulations during analysis.

1.5 Mass spectrometry instrumentation

This section presents the schematics of the two mass spectrometers used in this dissertation: the LTQ Orbitrap XL and the QExactive. Both instruments are considered “hybrid” mass spectrometers because multiple types of mass analyzers are utilized. The LTQ Orbitrap XL combines a linear ion trap, quadrupole, and Orbitrap mass analyzer (Figure 1.3), as shown in Figure 1.4. The QExactive combines a quadrupole with the Orbitrap mass analyzer as shown in Figure 1.5. Complex ion optics, electronics, and vacuum pumps are essential in the instrumentation, but their discussion is beyond the scope of this dissertation. The Orbitrap mass analyzer has been essential in advancing proteomics because it allows high mass accuracy and resolution comparable to FT-CIR mass analyzers, but without the prohibitive operation costs or slow cycle times.

1.5.1 The Orbitrap mass analyzer

In both the QExactive and LTQ Orbitrap XL, the ionized analytes are transmitted through the upstream mass filters and ion optics and accumulated in a radio frequency (RF)-only curved axis quadrupole, commonly referred to as the C-trap. Once the specified number of ions is collected according to the automatic gain control (AGC) setting, the entire packet is accelerated to high kinetic energy and injected tangentially into the Orbitrap. The ions are trapped in orbit around the inner electrode by an electrostatic field. The ion packet is introduced at an offset to the equator of the inner electrode resulting in harmonic oscillation along the z-axis of the central electrode at frequencies inversely proportional to the square root of each ion’s m/z. The image current is measured by the outer electrodes and the signal is Fourier transformed to produce a high resolution spectrum. This process is illustrated in Figure 1.3 reproduced from Eliuk et. al (2015). The resolution of the spectrum is proportional to the time over which
the image current signal, also referred to as the transient, is acquired and diminishes with the square root of the m/z. While the QExactive and LTQ Orbitrap XL use essentially the same mass analyzer, the QExactive resolution is improved using an enhanced Fourier Transform (eFT) algorithm. The original Orbitrap mass analyzer as used in the QExactive achieves a resolution of 200,000 at 400 m/z by acquiring a transient for 1024 milliseconds. The miniaturized high-field Orbitrap mass analyzer (not used in this dissertation) reaches a resolution of 240,000 at 400 m/z from a 768 millisecond transient. The transient length is proportional to the resolution, so scan speed can be doubled each time the resolution is reduced by a factor of two, but a high resolution is often necessary for unambiguous analyte identification.

Figure 1.3 – C-trap and Orbitrap Illustration. An Illustration of the C-trap and Orbitrap transient acquisition reproduced with permission from Eliuk et al. (2015).
1.5.2 LTQ Orbitrap XL

In this dissertation, the LTQ Orbitrap XL was operated in the high-low resolution mode, meaning that the MS1 scan was acquired at high resolution in the Orbitrap mass analyzer, while the MS2 scans were simultaneously acquired at low resolution in the linear ion trap. The linear ion trap scans much faster than the Orbitrap, but it is limited to roughly 0.5 Da resolution. The ions are trapped in the linear ion trap up to a set capacity using a quadrupole RF field and a compensating DC voltage. Ion populations are then selectively excited and ejected tangentially to strike an electron multiplier and produce a signal proportional to the number of ions in the ejected m/z population. Through routine calibration, the ion trap field and potentials are associated with ejected m/z values, and the signal from the electron multiplier is interpreted as a low resolution spectrum. During a DDA MS2 scan, a narrow (e.g. 2 Th) m/z range corresponding to the intended precursor is transmitted by the upstream quadrupole mass filter while all other m/z values are filtered out. The ions are accumulated in the ion trap up to a set capacity or for a maximum ion injection time. Then, an inert gas (typically helium) is injected, and the collected ions are destabilized through collisions resulting in cleavage of the weakest bonds in a process known as collision induced dissociation (CID).

Figure 1.4 – LTQ Orbitrap XL Schematic. Schematic of the LTQ Orbitrap XL with the mass analyzers labeled, the background is reproduced with permission from Makarov et al. (2006).
1.5.3 QExactive and QExactive Plus

The data presented in Chapters 3 and 4 were acquired using a QExactive Plus and a QExactive mass spectrometer, respectively. The QExactive Plus is identical to the QExactive model except for an improved quadrupole which is more efficient in transmitting the selected m/z range and a bent flatopole which increases robustness. Figure 1.4 shows a schematic of the QExactive Plus with the key components labeled. Contrary to the LTQ Orbitrap XL, both the fragment and precursor scans are acquired at high resolution in the Orbitrap. While the LTQ Orbitrap XL can acquire MS1 and MS2 scans in parallel, the QExactive only has one mass analyzer, and scans are acquired sequentially. Regardless of the data acquisition strategy, the QExactive is operated most efficiently when the transient time, which dictates scan resolution, is matched to the fill time of the C-trap. In this scheme, a population of ions is transmitted to the Orbitrap, and while the scan is acquired, the next pack of ions is collected in the C-trap. This second pack is immediately injected into the Orbitrap when the previous scan has been completed, allowing continuous Orbitrap operation.\textsuperscript{45}

The QExactive quadrupole uses an RF-field and a DC voltage to selectively transmit ions within a specified m/z range; all other m/z ranges are destabilized and are filtered out. To acquire an MS1 scan, the relevant m/z range (e.g. 400-1,600 m/z) is transmitted by the quardupole. To acquire the MS2 scan, the quadrupole selects the relevant m/z range, and the ions pass through the C-trap into the high energy collision dissociation (HCD) cell. Here, the ions are focused into a beam and collided with an inert gas (typically nitrogen) at a high energy resulting in rapid decomposition. The fragment ions are then sent back into the C-trap and transmitted into the Orbitrap. Both CID and HCD fragmentation depend on dissociation of the weakest bonds through collisional activation. Thus it is expected that similar fragment ions are produced by both, but due to difference in energies the generated spectra may not be identical.\textsuperscript{53} In DDA analysis the precursor ions selected from each MS1 scan are transmitted sequentially and individually in narrow (e.g. 2 m/z) isolation windows. In DIA analysis, ranges of m/z values are transmitted sequentially to scan through the entire relevant m/z range. The QExactive instruments are more sensitive than the LTQ Orbitrap XL due to improved ion transmission and have an increased resolution due to implementation of the enhanced Fourier Transform algorithm. Furthermore, the QExactive is able to
acquire up to 12 high resolution fragment scans per second, compared to only 3-4 scans/second in the LTQ Orbitrap XL.\textsuperscript{45, 48}

![Diagram of QExactive Plus schematic](image)

**Figure 1.5 – QExactive Plus schematic.**\textsuperscript{45}
Schematic of the QExactive Plus with the mass analyzers labeled, the figure is reproduced with permission from Michalski et. al. (2011).

1.6 LC-MS based protein analysis

1.6.1 Proteins as substrates for LC-MS analysis

Proteins are linear polymers of amino acid monomers linked in sequences specified by the encoding genes. Despite the challenges described in subsequent sections, proteins have significant benefits as substrates of LC-MS analysis. They are assembled from 20 monomers of which only 2 are isobaric (leucine and isoleucine). The monomers are L-chiral with the exception of achiral glycine, and the peptide bonds connecting amino acid residues are predictably cleavable using common mass spectrometry fragmentation techniques. Furthermore, protein sequences can be translated from sequenced genomes of the source organism which limits the $20^n$ (n - number of amino acid residues in the protein sequence) possible permutations to only the $\sim 10^4$ sequences encoded in the organism’s genome, thus greatly reducing the possible interpretations of LC-MS data.\textsuperscript{54}
However, the gene to protein to phenotype model is, in reality, an oversimplification. A single gene can be transcribed into multiple RNA splice variants that could produce protein sequences not directly predictable from the organism genome.\textsuperscript{55} An even more important complication in LC-MS protein analysis is the presence of post-translational modifications (PTMs). PTMs are covalent, and generally enzyme-driven, chemical modifications of protein residues (amino acids incorporated into a polymer chain). Some modifications are relatively simple like phosphorylation, where the hydroxyl active group of tyrosine, threonine, serine, or histidine is modified to a phosphate.\textsuperscript{56} Other modifications like glycosylation\textsuperscript{57} and ubiquitinylation\textsuperscript{58} involve a branched polymer moiety which generates a myriad of structural permutations. There are hundreds of known PTMs, some of which are known to modulate protein activity, while others are not well understood.\textsuperscript{59} PTMs can be detected by LC-MS as mass-shifts in the precursor mass and localized using the observed mass-shifts in the MS2 fragment ions. Most PTMs are not predictable from the genome, which complicates data analysis (discussed further in Section 1.6).\textsuperscript{60} Thus, while the polymeric structure of proteins is advantageous in LC-MS analysis there are several challenges which are discussed throughout this dissertation.

1.6.2 Applications of LC-MS based proteomic analysis

Global protein analysis, also referred to as proteomics,\textsuperscript{4, 61} is a vital modality for investigation of cellular behavior and has become an integral component of Systems Biology, the holistic study of the interactions of biological components.\textsuperscript{62} Prior to the maturation of LC-MS based proteomic analysis, Systems Biology relied on genome sequencing, or genomics, and microarray based RNA analysis, or transcriptomics, to investigate cellular behavior. The rates of gene transcription were used as indicators of cellular activity.\textsuperscript{63} However, cellular behavior is regulated by protein activity which correlates with abundance but is also controlled by PTMs, localization, and the flux of metabolites.\textsuperscript{64} While gene transcription can provide some indication about protein levels, it does not account for protein degradation or translation rates. Protein's role in biological systems has previously been investigated using knock-out studies, SDS-PAGE, gel electrophoresis, western blotting, ELISA, and ELISA arrays.\textsuperscript{65} However, these techniques examine only a handful of proteins, and modern LC-MS based methods can routinely identify and quantitate thousands of proteins in a matter of hours, allowing a comprehensive survey of protein
species that closely parallels the throughput of modern transcriptomic and genomic investigations. As discussed in the previous section, LC-MS can detect PTMs, allowing an even deeper understanding of protein activity. Metabolite flux, also an important factor in protein activity and cellular behavior, can be analyzed with high-throughput metabolite analysis, or metabolomics, by LC-MS analysis. Investigations combining analysis of gene transcription, protein levels and modifications, and metabolites are referred to as “multi-omic” and can be used to study complex biological networks.

Another important application of LC-MS based protein analysis is the characterization of therapeutic protein products. According to DrugBank.ca, there are 340 protein based therapeutics approved in the United States in 2016, comprising a roughly $140 billion market. Unlike small molecule drugs, in which the active ingredient is a single molecular entity, protein therapeutics contain a mixture of protein variants along with process-related impurities. Cell cultures can produce proteins with a range of PTMs, splice variants, and truncations. All of these forms have to be identified and quantitated to ensure that the produced therapeutic is safe and effective. Furthermore, endogenous cell proteins can be co-purified with the therapeutic protein and present an important class of impurities referred to as host-cell proteins (HCP). HCPs are potentially immunogenic and thus must be accurately identified and quantitated, as discussed in Chapter 3. LC-MS based proteomic analysis can sequence the various therapeutic variants as well as identify and quantitate HCP species, providing an orthogonal analytical method to conventionally used techniques like peptide mapping by LC with ultraviolet absorption (UV) detection, ELISA, and polyacrylamide gel based techniques.

1.6.3 Bottom-up proteomics

LC-MS based proteomics is most frequently carried out in a bottom-up scheme, where the proteins are enzymatically or chemically digested into smaller sequences called peptides. The advantage of this approach is that most peptides readily ionize, are efficiently separated by liquid chromatography, and are within the ideal size range for detection and fragmentation by most mass analyzers. Unique peptide sequences (only one gene in the genome translates to this sequence) are used as surrogate markers for the proteins from which they were derived. Bottom-up proteomics measures the translation of the
identified genes, with the ability to identify and quantitate PTMs, but does not provide information about
the specific protein variants translated from the gene. Analysis of intact proteins, referred to as top-down
analysis, can provide information about the levels of various protein isoforms and the inter-play of post-
translational modifications, but suffers from challenges in separation, ionization and fragmentation of
intact protein species, and upper limits of m/z range of MS instrumentation. While the field of top-down
proteomics is continually improving, this dissertation is focused on the currently dominant bottom-up
strategy with potential in translating the presented findings to top-down analysis.

There are several reliable reagents for protein digestion including cyanogen bromide, chymotrypsin,
and Endoproteinase Lys-C, but trypsin is most widely used. Trypsin is a protease which efficiently, and
with high fidelity, cleaves denatured proteins at the carboxyl end of each lysine and arginine residue that
is not followed by proline. The peptides produced through trypsin digestion contain enough residues (6-20)
to uniquely identify the source protein and ionize in a relatively narrow m/z range (400-1600 m/z). As
an added benefit, tryptic peptides contain an N-terminus along with at least one lysine or arginine residue
which results in +2 or greater charge (for most peptides) during ionization at low pH and allows the mass
spectrometer to ignore +1 charged non-peptide contaminants during selection of MS2 candidates. An
unfavorable caveat of digestion is that the complexity of the mixture is amplified as each individual protein
generates tens to hundreds of tryptic peptides and a digested proteome yields hundreds of thousands of
peptides. Thus separation of these digested samples requires high resolution and peak capacity. While
HILIC provides a viable alternative, this separation is most frequently achieved by reversed phase LC.

1.6.4 Reversed phase LC in bottom-up proteomics

Reversed phase separation at low pH (e.g. 2.6-2.8 pH with 0.1% formic acid modifier) using low
micron diameter fully porous or solid core porous particles covalently coated with 18 carbon alkyl chains
(C18) is the most frequently used LC mode in “bottom-up” proteomics. Peptides are readily retained on
the hydrophobic surface in an aqueous mobile phase and are released when the organic content of the
mobile phase reaches a threshold proportional to each peptide’s hydrophobicity. Peptide hydrophobicity
is affected by both the polarity of the peptide residues and their sequence, enabling the separation of
isobaric peptides (same residues but in a different sequence), which cannot be distinguished by MS1 alone. Reversed phase retention follows an on-off mechanism, meaning that retention is high when the mobile phase is insufficiently organic, and is negligible above the threshold. Thus, reversed phase separation produces sharp peaks and achieves high resolution and peak capacity as long as the organic content of the mobile phase can be gradually increased in a gradient. The gradient is achieved by mixing the flow from two pumps delivering aqueous and organic (most commonly acetonitrile) solvents. The flow-rate of the organic solvent pump is gradually increased while the flow-rate of the aqueous solvent pump is gradually decreased. The peak capacity ($P_c$) of a gradient separation is proportional to the effective gradient time ($T_g$ – the difference in elution time of the last eluting peptide and the first eluting peptide) divided by the average chromatographic peak width ($W$): \(^{23}\)

\[
P_c = 1 + \frac{T_g}{W} \quad \text{(Equation 4)}
\]

Reversed phase peaks broaden ($W$ increases) with longer gradient times, resulting in limited increase in peak capacity with prolonged separations. MS-based analysis of complex biological samples can often require more resolution and peak capacity than can be accomplished by conventional reversed phase separation. The efficiency of LC has been increased by reduced stationary phase particle diameter, novel stationary phase geometries (e.g. monolithic), and by combining multiple chromatographic separations (2D-LC).

The Van Deemter equation (Section 1.2) models the benefits of reduced bead diameter: (1) improved uniformity of flow due to more homogeneous packing and (2) reducing band broadening due to diffusion into the stagnant mobile phase inside the particle pores. These effects result in lower theoretical plate height, higher resolution and peak capacity, but also increase backpressure. As such, higher pressure pumps have been developed to operate columns containing sub 2 micron diameter particles under a high pressure (over 600 bar). This technology, referred to as ultra-high performance LC (UPLC), has been instrumental in advancing LC-MS based analysis.\(^7\) A comparison of conventional high performance LC (HPLC) and UPLC reversed phase separations is presented in Figure 1.6.
Figure 1.6 – Increased peak capacity with UPLC.\textsuperscript{78}
Two identical separations using columns packed with 1.9 micron (bottom) and 3 micron (top) diameter beads. The UPLC separation achieves more than double peak capacity over the same separation due to reduced band broadening. Reproduced with permission from Swartz \textit{et. al.} (2005).

Alternatively, LC efficiency is improved with novel solid phase modalities. Porous monolithic columns are gaining ground as an excellent alternative to UPLC. The high porosity of the monolith results in a very high surface area to volume ratio, leading to high loading capacity (amount of sample that can be adequately separated on the column). Furthermore monolithic columns are very permeable, resulting in low backpressure which allows the use of potentially long and narrow columns.\textsuperscript{79, 80} Other modalities like pellicular particles\textsuperscript{81} and solid core particles\textsuperscript{82} also offer superior resolution compared to conventional microbeads by reducing band broadening due to pore diffusion with shallow pore depth and consequently low stagnant mobile phase volume. However, one important limitation of pairing high resolution LC to MS is that the mass spectrometer sampling rate must be sufficient to obtain MS2 information for all eluting peptides before their elution is finished. Thus, the benefits of ever increasing chromatographic resolution
are limited by the performance of the MS. Fortunately, peak capacity can also be increased by combining the separation of multiple LC methods (2D-LC).

Complex samples can be separated into several less complex fractions, and each fraction can then be analyzed as an individual sample. In proteomics it is possible to fractionate the sample at the protein or the peptide level. In GeLC-MS, the intact proteins are separated on a polyacrylamide gel either in the native or denatured state. Each gel bands is then excised, subjected to the digestion protocol, and the extracted peptides are then analyzed by LC-MS as separate samples. Fractionation can also be carried out at the peptide level through multiple LC separations. The initial LC separation can fractionate peptides based on one retention mechanism, and the second, an orthogonal LC separation method, is paired to MS to analyze each fraction. For example, ion exchange chromatography, which separates analytes based on charge at the mobile phase pH is orthogonal to reversed phase separation based on hydrophobicity. Combining chromatographic fractionation with LC-MS analysis is referred to as 2D-LC-MS (or 3D-LC-MS if two sets of fractionating separations are used). In Chapter 3 a 2D-LC separation is carried out by first separating peptides by reversed phase at a high pH and then carrying out the analysis using low pH reversed phase LC-MS. The reduction in charge of the peptide at a high pH has a varied effect on peptide hydrophobicity, which results in partially orthogonal separations between low and high pH reversed phase. Figure 1.7 presents an illustration of this separation which distributes the high intensity peptides from the most abundant protein over multiple fractions thus reducing the masking of less abundant species.

With high resolution LC and modern MS instruments, which acquire tens of spectra per second, a typical LC-MS run produces tens of thousands of spectra. Manual analysis of these spectra is unfeasible so proteomics data is processed by automated algorithms, which are introduced in the following section.
Figure 1.7 – High - low pH reversed phase LC separation.
In High-low pH reversed phase separation the peptides are first separated at a high pH (~10) (top) and each collected fractions is analyzed by low pH (~2.7) reversed phase LC coupled to MS (bottom). This example shows the ion chromatogram and the ion map demonstrating improved sensitivity where the low intensity species are separated from the high intensity species.

1.7 Bottom-up proteomic data analysis

Thousands of spectra are acquired over the course of an LC-MS experiment and processing by automated means is required. Data from each data acquisition strategy require a different approach. DDA data are analogous to a catalogue of photographs, with each spectrum representing an image from which the identity of one analyte can be determined. DIA data are analogous to a film where the elution of fragment and precursor ions is recorded over time, and the identities of individual analytes can be inferred by tracking the overlap of fragments and precursors across their elution peaks. These concepts are
Illustrated in Figure 1.8. Targeted MS data have a time element like DIA, but the spectra are focused on individual analytes like DDA. This section will first discuss analysis of DDA data, followed by analysis of DIA data which implements a strategy used in processing of targeted MS data. The section is concluded with a summary of the general challenges of interpreting bottom-up proteomic data.

**Figure 1.8 – Illustration of DDA and DIA data.**
DDA data isolates individual ions (top left) to generate spectra for individual peptides (bottom left). DIA repeatedly fragments large isolation windows (top right) allowing tracing of fragment ion elution (bottom left). The figure is reproduced from Egertson et al. (2015).

### 1.7.1 DDA data analysis

As mentioned in Section 1.5, protein analysis by LC-MS is possible with the availability of a protein sequence database translated from the organism genome or transcriptome. This database can be processed in silico, using rules matching the specificity of the protease used in sample preparation. For example, a list of tryptic peptides (peptides generated from digesting proteins with trypsin) is generated by cleaving each protein sequence after every lysine and arginine residue that is not followed by proline.
(which is known to interfere with trypsin activity). The acquired DDA data can then be matched to these theoretical peptides and then to the protein from which the peptides were derived using one of three fundamental approaches: *de novo* (e.g. PEAKS,\textsuperscript{87} Novor),\textsuperscript{88} database searching (e.g. Sequest,\textsuperscript{89} MASCOT,\textsuperscript{90} MS Amanda,\textsuperscript{91} Andromeda,\textsuperscript{92} Myrimatch,\textsuperscript{93} and MS-GF+),\textsuperscript{94} and spectral matching (e.g. SpectraST).\textsuperscript{95}

In the *de novo* strategy, each spectrum is treated like a puzzle with the fragment ions interpreted as linear combinations of amino acid monomers called mass tags. *De novo* algorithms reconstruct the peptide sequence from the detected mass tags using the precursor ion m/z to limit possible interpretations (the mass tags must add up to the intact mass). The protein sequence database can then be used to help interpret spectra potentially matching multiple peptide sequences.\textsuperscript{96, 97} The database search strategy matches theoretical peptide fragments from the *in silico* database to the acquired spectra. Typically, a correlation score between the theoretical fragments of peptides in the database and the acquired spectra is used to generate potential matches.\textsuperscript{98} The spectral matching strategy is computationally efficient because the acquired spectra are compared to an assembled library of high quality spectra corresponding to known peptides. The acquired and reference spectra are represented as high dimensionality vectors with the magnitude of each term corresponding to the intensity within the corresponding m/z bin. Computation of the dot product of the reference and acquired vectors provides a computationally rapid metric of their similarity. A high dot product (close to 1) indicates a match between the acquired spectrum and one of the spectral library entries, and is interpreted as detection of the corresponding peptide. Spectral searching can be more specific than the *de novo* and database search strategies, because the relative intensities of the fragment ions are also taken into consideration in the matching process. However spectral matching can only identify peptides and proteins represented in the reference spectral library, and any errors, such as incorrect peptide-spectrum matches incorporated into the spectral library, may be propagated into the search results.\textsuperscript{95, 99}
1.7.2 Evaluation of search result confidence by target-decoy

Manual verification of all identifications is unfeasible, and automated approaches to evaluate the reliability of the identifications are implemented. Regardless of the data acquisition or processing strategy, the search identifications can be divided into four categories: true positive (TP) – identified and present in the sample, false positive (FP) – identified, but not present in the sample, true negative (TN) – identified and not present in the sample, and false negative (FN) – not identified but present in the sample. The search results represent the combined true and false positive populations, while the remaining theoretical peptides that were not identified in the search represent the true and false negative populations.

Every search strategy strives to maximize TP while minimizing FP and FN. This is accomplished by scoring each peptide-spectrum match (PSM) based on several quality attributes, the most important of which are the mass errors (the difference between the theoretical mass and the observed mass) of the precursor and the fragments, and the number and ratio (number of matched ions divided by the total number of ions in the spectrum) of matched fragment ions. The scoring scheme is intended to score TP matches favorably and FP matches poorly. The efficiency of this process is tested by adding decoy sequences (sequences of artificial peptides which are not present in the sample) into the database. If the decoy population is not qualitatively different than the database peptides (i.e. same length, residues distribution, and similar number of sequences), then the number of decoys matched by the search algorithm at a given score threshold should mimic the number of FP peptide matches at that threshold. One simple approach to generating the decoy population is through reversal of the protein sequences in the database and \textit{in silico} digestion. Figure 1.9 shows the decoy and target (non-decoy) peptide score distributions generated by the \textit{de novo} PEAKS search engine. In the figure, a score threshold is selected to incorporate 1% decoy sequences, which is equivalent to accepting 1% false positive peptide identifications or, more commonly referred to as 1% FDR. An important caveat of this approach for evaluation of identification confidence is that the TP population has to be roughly the same size as the decoy population; the model breaks down if the TP population is too small as discussed in Chapter 3.
Figure 1.9 – Typical score distribution of decoy and target peptide spectral matches.
Output from a de novo search in the PEAKS 7.0 suite. The histogram presents the score distribution of decoy (red) and target (blue) matches. The decoy population is low scoring and the overlapping target matches below the dashed vertical line (the FDR threshold) are not accepted. Matches above are threshold accepted with the FDR estimated from the number of decoy matches above the threshold.

1.7.3 DIA data analysis

The earliest variation of DIA, MS\textsuperscript{E}, was carried out on Waters instruments by alternation of a precursor, MS1 scan, over the relevant m/z range, with a scan in which all precursors in the range were fragmented together.\textsuperscript{103} MS\textsuperscript{E} is the simplest data acquisition strategy that allows chromatographic tracking of all precursors and their fragments. MS\textsuperscript{E} data were analyzed by evaluating the correlation of the elution profiles of precursor and fragment ions and then grouping precursors and fragment ions with matching elution profiles. The matched precursors and fragments were searched against the full protein sequence database to identify peptides using proprietary algorithms similar to those described for DDA data analysis. Recently, this DIA data analysis strategy became available for all instruments, and DIA variations (SWATH,\textsuperscript{40, 41} MSX-DIA,\textsuperscript{42} and pSMART,\textsuperscript{43} see Table 1.2) through DIA-Umpire.\textsuperscript{104} DIA-Umpire groups precursors and fragments based on Pearson correlation and generates pseudo-DDA spectra. These pseudo-DDA spectra can then be searched using any DDA processing algorithm against a
database of relevant protein sequences. In this dissertation, this strategy is referred to as untargeted DIA data analysis because the search is not limited to a library of assays, as with the targeted strategy described in the subsequent paragraph.

DIA data can also be analyzed using the peptide-centric, or targeted, strategy. This approach originated from analysis of targeted MS data. In targeted MS, the concentration of an analyte is determined by the integration of characteristic fragment ion chromatograms generated when the analyte is isolated and fragmented throughout the course of its elution. DIA data contain repeated measurements of fragments from all precursors over the course of their elution. Thus, the same strategy can be implemented for detection and quantitation of analytes provided that the characteristic fragment ions and their retention times are known. This required information is referred to as an assay in the context of this dissertation. A collection of assays is referred to as an assay library, and large assay libraries are derived from spectral libraries, collections of reference spectra and normalized retention times, by extracting the annotated transitions (usually high intensity b and y ions). Spectral libraries are most efficiently generated from DDA analysis of the sample or related samples, because thousands of quality spectra which identify peptides with high confidence can be acquired in a single run.

In the targeted strategy, DIA data can be interrogated by thousands of assays, and a false discovery rate (FDR) model similar to the model described for DDA can be applied by adding an equal population of decoy mass-shifted assays and by scoring each potential peak match using several attributes including mass accuracy, co-elution, peak shape, and dot product. A single peak score is derived from a weighted combination of the peak sub-scores with the weights determined by a machine learning algorithm such as mProphet to maximize the discrimination between the decoy and target populations. Then, a threshold for the final score is set to accept the number of decoys corresponding to the desired FDR. This approach essentially transforms the data analysis in to a detection problem; instead of matching the data to an unknown peptide derived from the protein sequence database, the search strategy tests the confidence in the hypothesis that a given peptide in the assay library is present in the sample. Hence, the strategy is referred to as peptide-centric.
In both DDA and DIA analyses, accurate high resolution measurement of the intact precursor mass is essential to increase the specificity of identification. The higher the accuracy and resolution of the measurement, the fewer theoretical peptides can possibly match the data. High accuracy and resolution measurement of the fragment masses can provide even more specific identifications, and this principle is highlighted in Chapter 3. Despite the high resolution and mass accuracy available with modern MS instruments only roughly a third of the acquired spectra are identified in DDA analysis, and the subsequent section describes some of the challenges of processing bottom-up data.

1.7.4 Complications in bottom-up data analysis

Proteomic data analysis is complicated by the presence of PTMs and by incomplete, or non-specific enzyme digestion. A PTM which affects one specific residue doubles the number of peptide permutations which contain that residue, and this complexity is amplified by peptides which contain multiple modifiable residues. When multiple PTMs are considered, the number of possible permutations explodes. Furthermore, the digestion procedure is not perfect and may generate incompletely digested peptides, or peptides that are truncated by a nonspecific cleavage. These complications lead to a much higher number of possible peptide permutations than predicted from the protein sequence database. While data analysis using a search space (possible peptide permutations) expanded beyond the database of protein sequences derived from an organism genome by several orders of magnitude is feasible with modern computers, regulation of the false discovery rate can be problematic. Since each PTM and non-specific cleavage multiply the population of potential peptide matches, they also multiply the decoy population. Since the population of peptides in the sample is static, a disproportionate increase of the decoy population will result in an inflated FDR, because more decoy sequences will sporadically score over any set threshold. Efforts can be made to control the decoy population, but typically only 1-3 of variable modifications are used in data searches. This is an imperfect solution that results in unidentified high quality spectra and missed modified peptides.

A further challenge in bottom-up data analysis is inference of proteins from the identified peptides. Some peptides are unique to a single protein sequence in the entire database and thus produce a reliable
identification of that protein. However, more frequently, the peptide sequence is homologous between multiple proteins, and even more frequently homologous between protein variants. Homologous peptides are impossible to unambiguously attribute to any specific protein variant. One of the most common approaches to this problem is to group proteins based on homologous peptides, and then state that one or more entries in the protein group have been detected. Another approach used in protein inference is the parsimony principle, which attempts to explain the identified peptides with as few protein groups as possible. When the parsimony principle is implemented, a peptide which can be attributed to multiple proteins is attributed to the group with the most identifying peptides, and a minimal number of protein groups is used to encompass all of the identified peptides.\textsuperscript{110} Recently, the use of Bayesian statistics has become available to provide a statistical confidence to protein inference.\textsuperscript{111} Statistical basis for protein inference is an important step, but it only moderates the fundamental limitation of bottom-up proteomics, which is that once the proteins are digested, it is impossible to attribute a peptide to a specific proteoform (a protein sequence which could be modified post-translation).\textsuperscript{112} However, the ability to state that a protein in some form(s) or variant(s) is present in the sample for thousands of proteins in each experiment with high confidence is still incredibly valuable. But even more valuable is the ability to measure the concentration of the identified protein groups.

1.8 Quantitative LC-MS proteomics

Modern LC-MS proteomics recognizes that identification of the protein repertoire associated with a given sample is insufficient and that quantitative analysis is required to advance the understanding of biological systems or to monitor the quality of biopharmaceuticals.\textsuperscript{113} In Systems Biology, different system states manifest as perturbations in protein concentrations and complete elimination of a given protein set is relatively rare. Furthermore, proteins may not be detected due to insufficient analytical sensitivity, and the failure to detect protein species does not indicate absence. Thus, meaningful conclusions about system behavior require quantitation. In biopharmaceutical analysis, the effects of different therapeutic protein variants, host cell proteins (HCPs), and other impurities are dose dependent. As such, the components must be quantitated to accurately assess risk and ensure that the dosages of the active therapeutic variants are consistent. This section describes the quantitative LC-MS strategies, which can
be sorted into three categories: stable isotope labeled, label-free, and absolute quantitation using synthetic standards. Brief descriptions and compatibilities with different data acquisition strategies are presented in Table 1.3.

<table>
<thead>
<tr>
<th>Description</th>
<th>Compatibility with data acquisition strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heavy Isotope Labeled</strong></td>
<td>DDA</td>
</tr>
<tr>
<td>Metabolic labeling (SILAC)</td>
<td>Stable isotopes are metabolically incorporated into proteins using media containing labeled amino acids.</td>
</tr>
<tr>
<td>Isotopic tags ((^{18})O, ICAT)</td>
<td>Peptides or proteins are distinguished by mass due to the incorporation of heavy isotopes.</td>
</tr>
<tr>
<td>Isobaric tags (TMT, iTRAQ)</td>
<td>Peptides are labeled using a set of reagents of the same mass but different isotope distributions.</td>
</tr>
<tr>
<td><strong>Label-free</strong></td>
<td></td>
</tr>
<tr>
<td>MS1 peak area integration</td>
<td>Integration of precursor peak area.</td>
</tr>
<tr>
<td>MS2 peak area integration</td>
<td>Integration of peak area of characteristic fragments</td>
</tr>
<tr>
<td>Spectral counting</td>
<td>The number of PSMs corresponding to a protein correlates with protein abundance.</td>
</tr>
<tr>
<td><strong>Isotope dilution MS</strong></td>
<td></td>
</tr>
<tr>
<td>Comparison to SIL standards</td>
<td>Peptides homologues synthesized with heavy isotopes are spiked-in to quantitate the endogenous peptides based on peak area ratios.</td>
</tr>
</tbody>
</table>

Table 1.3 – Comparison of quantitative proteomic strategies.

1.8.1 Stable isotope labeled quantitation

The labeled strategy is typically used for comparative multiplexed quantitation, where the reagents are used to differentially label peptides from multiple samples, the samples are mixed, analyzed by LC-MS as a single sample, and the relative quantities of the peptides from each sample can be determined based on the reagent signal intensity. The labels are distinguishable by MS due to different incorporation of stable isotopes (\(^{13}\)C, \(^{15}\)N, \(^{18}\)O), but uniformly affect peptide chemical properties across samples.
preventing any biased shift in LC retention or ionization efficiency. Analytes can be labeled chemically, enzymatically, or metabolically, and the labeling reagents are either isotopic (distinguishable by intact mass) or isobaric (same precursor mass but different fragment masses).

The simplest variation of labeled quantitation is $^{18}\text{O}$ labeling during proteolysis. Trypsin$^{115}$ Glu-C$^{116}$ and other proteases$^{117}$ incorporate oxygen from water during hydrolysis of the amide bond. Thus one sample is digested under normal conditions and the second sample is digested in $^{18}\text{O}$ water, the samples are mixed together, and analyzed by LC-MS. The peptides from the second sample are 2 Da heavier than their counterparts and the peak area ratio for each pair reflect the peptide fold-change between the two samples. While this approach is simple to implement, it is limited to comparison of 2 samples and quantitative accuracy may be lost due to back exchange of $^{18}\text{O}$ to $^{16}\text{O}$. Isotopic labels can also be incorporated metabolically by growing cell cultures in media containing isotopically labeled amino acids in SILAC (stable isotope labeled amino acids in cell culture) experiments. The amino acids are metabolically incorporated into the cellular proteins providing a label for downstream LC-MS analysis. In SILAC the variety of reagents allows a higher multiplex (number of samples which can be compared in a single experiment) than $^{18}\text{O}$, and the proteins from all samples can be digested together, thus eliminating this source of variability and allowing accurate comparative quantitation.$^{118}$ Unfortunately SILAC requires metabolic incorporation of the reagents which limits the applications to cell culture samples. Isotopic labels can also be incorporated by chemical means, for example ICAT (isotope-coded affinity tag) reacts specifically with cysteine sulfhydryl groups and contains 9 $^{13}\text{C}$ atoms in the heavy tag resulting in a 9 Da difference.$^{119}$

Isobaric mass tags have the same chemical composition and same intact mass but the isotopes are distributed in such a way that fragmentation produces different reporter ions for each reagent, and the relative intensities of these reporter ions can be used for comparative quantitation of the labeled peptides (Figure 1.9). There are two commercially available sets of reagents TMT$^{120}$ and iTRAQ$^{121}$ Both react specifically with primary amine groups (lysine side chain and N-termini) of peptides. iTRAQ reagents are available for 4-plex and 8-plex analysis, while TMT reagents can be used for 2, 6, and, with high mass resolution, 10 - plex$^{122}$ analysis. Quantitation based on MS2 should potentially have higher signal to noise
than MS1 based quantitation because precursors are isolated from other co-eluting ions. However isolation of individual precursors is required to prevent convolution of reporter ion signals from multiple precursors. As such, DIA is incompatible with isobaric mass tag quantitation, and fractionation (2D-LC) is often required to prevent co-elution of near-isobaric peptides for accurate quantitation in complex samples.123,124

Labeled quantitation offers better quantitative precision and accuracy than label-free quantitation and multiplexed analysis substantially increases throughput. On the other hand, label-free analysis requires minimal manipulation of the sample and results in broader proteome coverage and higher dynamic range.125 Furthermore, while label-free analysis is limited to one sample at a time, an unlimited number of samples can be compared.
Figure 1.10 – 6-plex TMT reagents and workflow.
Top left shows the functional groups of the TMT reagents, the top right shows the distributions of stable isotopes in each reagent differentiating the reporter ion masses. In the workflow each samples is labeled using a different reagent, the samples are mixed together, and analyzed together. The MS2 spectra for each peptide contain ions used for identification and reporter ions used for relative quantitation.
1.8.2 High throughput label-free quantitation

In label-free LC-MS the linear correlation between peak area and analyte concentration is exploited for quantitation. The peptide peak areas can be used to compare relative protein abundance within a sample, and between samples if the analytical parameters are replicated between runs. The principle obstacle in label-free quantitation is ion suppression.\textsuperscript{126} Ion suppression from co-eluting species can significantly reduce signal intensity leading to variance between samples and between replicate experiments due to shifts in LC retention. In DDA there are two options for label free quantitation: spectral counting and integration of precursor ion peak area. In Top N DDA the number of PSMs attributed to a protein has been shown to correlate with protein concentration and can be used for rough quantitation.\textsuperscript{127, 128} Quantitation based on precursor ion peak area (MS1) is more accurate and can be used to quantitate peptides which were not fragmented during a DDA run by matching high-resolution and high-accuracy mass and retention time to peptides identified in parallel experiments.\textsuperscript{129, 130} It is also important to note that MS1 peak area can be measured with all of the data acquisition strategies.

DIA can also (in addition to MS1 peak area) trace the peak areas for fragment ions and quantitate based on MS2 peak areas. Quantitation based on fragment peak area is potentially more specific than MS1 peak area because isobaric co-eluting peptides can be distinguished, this approach also offers the opportunity to quantitate peptides which are masked by high intensity co-eluting species.\textsuperscript{41} MS2 quantitation is also more sensitive because ions are accumulated through the instrument automatic gain control while other species are filtered out, resulting in higher signal to noise. This mechanism is even more efficiently implemented in targeted MS analysis where individual precursor ions are isolated and accumulated in corresponding narrow isolation windows.

1.8.3 Targeted label-free quantitation

As mentioned earlier, targeted analysis is most commonly used for accurate and precise quantitation of a small set of peptides. In proteomic studies, DDA or DIA experiments are used to identify and quantitate thousands of proteins and then the proteins of interest are quantitated by targeted analysis.\textsuperscript{131, 132} Targeted quantitative analysis was initially carried out on triple quadruple instruments using the
selected reaction monitoring (SRM) scheme, in which a narrow m/z range corresponding to the intended analyte was isolated using the first quadrupole, fragmented in the second quadrupole, and the ion current for a characteristic fragment ion isolated in the third quadrupole was measured over the elution of the analyte. SRM is linear down to low nanogram / milliliter levels even in complex matrixes, but requires considerable optimization for each analyte to select the most intense, and specific, fragment ion and optimize fragmentation parameters. Alternatively, instruments where all ions in the MS2 scan are detected simultaneously, like the QExactive, can implement a strategy referred to as parallel reaction monitoring (PRM). PRM achieves comparable performance to SRM, without optimization, because all fragment ions are monitored in parallel and the best ions can be selected for quantitation during data processing.

Regardless of data acquisition strategy, ionization suppression from high intensity co-eluting species can vary between samples and complicate comparison. Heavy isotope labeling can be used to overcome this source of variance. However because MS signal can vary between instruments the measurements by both quantitative strategies are comparative and do not translate to real-world measurements without the use of standards.

1.8.4 Accurate and absolute quantitation by isotope dilution MS

Standards can be spiked into LC-MS analysis to generate a calibration curve which correlates measured signal to concentration. The most accurate quantitation is accomplished with stable isotope labeled (SIL) homologues of the relevant analytes. The synthesized SIL standards contain heavy residues and are spiked in at the approximate concentration of the endogenous analyte. The retention, response factor, and any suppression effects are identical between the standard and the endogenous analyte and the ratio of their peak areas is used to determine the accurate concentration of the endogenous analyte. This strategy, referred to as isotope dilution MS, is best implemented with targeted MS using the fragment peak areas for comparison, but is also compatible with DIA. Importantly, this approach is used in quantitation of a set of analytes for which the standards were synthesized. Synthesis of SIL peptide standards is relatively inexpensive and the concentration of around one hundred peptides...
can be accurately measured in a single targeted LC-MS run. However these measurements don’t directly translate to accurate measurement of protein concentration due to potential for incomplete digestion and modification/degradation of the endogenous or SIL peptides. Criteria are used to select peptide standards which mitigate these risks. These criteria include: exclusion of residues which can be oxidized or deamidated, unique tryptic sequence, and high response factor.\textsuperscript{137} Artifacts associated with digestion can be compensated for by using SIL protein standards which are spiked in before sample digestion and are subjected to the same stresses as the endogenous protein. This approach can potentially provide absolute protein quantitation, but is currently prohibitively expensive.\textsuperscript{138}

1.9 Conclusion

This chapter has summarized the data categories that can be acquired during LC-MS based proteomic analysis and introduced strategies for translating these data into protein identifications and quantitative measurements. Acquisition of precursor m/z’s using repeated MS1 scans over the course of chromatographic separation generates invaluable data for identification of the components. All data acquisition strategies can be implemented with routine MS1 scans, thus the key consideration is how best to acquire fragmentation data to address the requirements of the experiment. Targeted MS specifies a relatively limited population of precursors to be fragmented over the course of their elution to generate high quality fragment peak traces for quantitation. DDA selects precursors on the fly and identifies a large but ultimately limited portion of the analytes by acquiring MS2 spectra for individual precursors. Conversely, DIA traces fragments throughout their elution, allowing matching to precursor ions and potentially identifying all detected species. However, the generated DIA data is inherently ambiguous, and sensitivity is reduced when fragments from low abundance precursors are masked by fragments from co-fragmented high abundance ions. Each data acquisition strategy presents advantages and limitations. The benefits of combination and intelligent implementation of these strategies are presented in the subsequent chapters.

Chapter 2, a formatted version of a published article titled "Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling", demonstrates how sophisticated precursor ion selection rules increase the sensitivity of bottom-up proteomic analysis.\textsuperscript{139} Chapter 3 is a
formatted version of a submitted article titled “Comprehensive HCP Profiling by Targeted and Untargeted Analysis of DIA Mass Spectrometry Data with PRM Verification” and demonstrates how dynamic range is increased through a combination of DIA and targeted MS analysis (parallel reaction monitoring). Chapter 4 is a combination of excerpts from a published review on extracellular vesicles (EV) titled “Mass-Spectrometry-Based Molecular Characterization of Extracellular Vesicles: Lipidomics and Proteomics” and a summary of a pilot study implementing a combination of DDA and DIA for analysis of EVs isolated by size exclusion chromatography from blood plasma. Together, the investigations highlight the potential for continual improvement of LC-MS analysis through implementation of novel data acquisition strategies.
1.10 References

15. Cuatrecasas, P., Citation Classic - Protein-Purification by Affinity-Chromatography - Derivatizations of Agarose and Polyacrylamide Beads. *Current Contents/Life Sciences* 1980, (22), 16-16.


Ong, S. E.; Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology* 2005, 1, (5), 252-262.


123. Ow, S. Y.; Salim, M.; Noirel, J.; Evans, C.; Rehman, I.; Wright, P. C., iTRAQ Underestimation in Simple and Complex Mixtures: "The Good, the Bad and the Ugly". *Journal of Proteome Research* 2009, 8, (11), 5347-5355.


Chapter 2: Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling

This chapter is a formatted version of "Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic profiling."¹ The dissertation author wrote the manuscript, participated in the design of the precursor ion selection algorithms, and conducted the published experiments. Dr. Michael Belov and William Danielson (Spectroglyph) developed the Smart MS² platform and programmed the implemented algorithms. Lev Levitsky, under the supervision of Dr. Mikhail Gorshokov, programmed the script which executed the integrated post-acquisition database search. The project was supervised by Dr. Alexander Ivanov and Professor Barry Karger.
2.1 Abstract

Conventional Top N data-dependent acquisition (DDA) LC-MS/MS analysis identifies only a limited fraction of all detectable precursors because the ion-sampling rate of contemporary mass spectrometers is insufficient to target each precursor in a complex sample. Top N DDA preferentially targets high abundance precursors with limited sampling of low abundance precursors and repeated analyses only marginally improve sample coverage due to redundant precursor sampling. In this work, advanced precursor ion selection algorithms were developed and applied in the bottom-up analysis of HeLa cell lysate to overcome the above deficiencies. Precursors fragmented in previous runs were efficiently excluded using an automatically aligned exclusion list, which reduced overlap of identified peptides to ~10% between replicates. Exclusion of previously fragmented high abundance peptides allowed deeper probing of the HeLa proteome over replicate LC-MS runs resulting in the identification of 29% more peptides beyond the saturation level achievable using conventional Top N DDA. The gain in peptide identifications using the developed approach translated to the identification of several hundred low abundance protein groups, which were not detected by conventional Top N DDA. Exclusion of only identified peptides compared to the exclusion of all previously fragmented precursors resulted in an increase of 1,000 (~10%) additional peptide identifications over four runs suggesting the potential for further improvement in the depth of proteomic profiling using advanced precursor ion selection algorithms.

2.2 Introduction

In a typical precursor-driven “Top N” LC-MS data-dependent acquisition (DDA) bottom-up experiment, precursor ions are automatically selected for fragmentation, without a priori knowledge, using simple rules. The precursor selection logic includes highest intensity priority, intensity thresholds, exclusion of specified m/z ranges (e.g. 445.12 Th, polysiloxane), dynamic exclusion, and charge state selection. The most important goals of precursor selection logic are noise filtering based on charge and intensity of ion species (e.g. z > +1 for tryptic peptides and intensity above a set threshold), and deeper sample probing using dynamic exclusion, which prevents redundant selection of a precursor for a set
period of time after a fragment (MS^2) scan has been acquired. Although commonly used, the conventional Top N data acquisition strategy is limited by stochastic and biased sampling,^4, 5^ fragmentation of precursors before and after the chromatographic peak maximum,^6^ co-fragmentation of near isobaric co-eluting precursors resulting in difficult to interpret chimeric spectra,^7^ and redundant selection of the same precursors between replicate analyses or within individual LC-MS experiments. In bottom-up proteomic analysis of complex samples, these deficiencies lead to the identification of only a minority of detectable peptides within individual experiments (typically < 30%).^8^ Additional peptides can be identified by repeated analysis, but due to sampling bias towards high-intensity features, high abundance peptides are redundantly identified while low abundance peptides are neglected. Thus, repeated analysis eventually leads to saturation where few new peptides are identified by additional technical replicates.^9^

To circumvent the deficiencies of conventional DDA, multiplexed and data-independent acquisition (DIA) strategies have been implemented in the bottom-up proteomic analysis.^10^ In multiplexed data acquisition, several precursors are simultaneously selected and fragmented to increase the rate of precursor sampling. In DIA experiments, fragment information is acquired for all eluting precursors also by parallel fragmentation of multiple precursors. DIA does not require precursor selection, but rather uses a systematic scheme for scanning the entire m/z range of interest (e.g. MS^E^,^11^ SWATH,^12, 13^ MSX-DIA,^14^ and pSMART^15^). A given peptide is more likely to be identified from a spectrum obtained by isolation in a corresponding narrow m/z window (e.g. 2 Th) such as in conventional DDA than from a multiplexed DDA or DIA spectrum (e.g. 20-25 Th isolation windows) due to challenges in interpreting multiplexed spectra and due to improved selectivity and sensitivity (for the individual precursor) in narrow isolation windows. Furthermore, DIA is incompatible with isobaric tag quantitation due to inter-mixing of reporter ion signals when multiple precursors are co-fragmented.^16, 17^ However, DIA can serve as a means for obtaining fragmentation spectra for peptides that would otherwise be neglected by conventional Top N analysis, which results in complementary coverage between the two strategies.^18^

Aside from improving physical separation of analytes by multi-dimensional and high-resolution liquid chromatography,^19^ the depth of DDA-based proteomic profiling can be enhanced with strategies for improved precursor selection. The most straightforward solution for sampling peptides missed by conventional DDA is targeted analysis (e.g. parallel reaction monitoring) to subject peptide-like precursors
unidentified in the initial LC-MS run(s) to scheduled fragmentation in subsequent replicate analyses. Several versions of this solution have been demonstrated to be advantageous over the conventional approach. One notable example, Post Analysis Data Acquisition (PAnDA), performs an automated database search after conventional DDA analysis to generate a list of precursor targets which were not identified in the conventional analysis, followed by targeted MS analysis to fragment these features. The authors reported a 30.9% gain (3,849 vs. 2,941) and 20.5% gain in peptide and protein identifications, respectively, in the bottom-up analysis of *C. elegans* proteome after 6 iterations. A related strategy implemented on an Orbitrap Elite combined targeted and untargeted analysis by fragmenting preset peptides, when they were detected and performing conventional Top 15 DDA simultaneously to identify other peptides in the sample in an untargeted manner. Although effective, the targeted strategy is dependent on the reproducible appearance of the precursors at the same retention time between runs. This is, unfortunately, not always the case for low-intensity peptides in complex mixtures, where ion suppression from high abundance species and chromatographic variability can mask the precursor in a somewhat stochastic manner. If a peptide is not detected in the initial survey, it will not be targeted, and if the elution time of the peptide changes during targeted analysis, it will likely not be detected.

An alternative to targeted analysis for increasing protein coverage in shotgun proteomics is scheduled exclusion of *m/z* ranges corresponding to previously identified peptides for set retention time intervals in previous gradient runs. Accurate Mass Exclusion-based DDA (AMEx), has been reported to increase the number of identified peptides by 26% (4,490 vs. 3,564) over conventional DDA by iteratively excluding peptides which were identified in previous LC-MS/MS runs (6 iterations). However, the inability to adjust for retention time shifts and the lack of on-the-fly deconvolution algorithms for both the charge and isotopic states of precursor ions have hampered the benefits of conventional inter-experiment exclusion (enabled through Xcalibur) in our hands. It is possible to extend the control of LTQ Orbitrap-based mass spectrometers (Thermo Fisher Scientific) beyond the standard functionality allotted by Xcalibur through LTQ COM Object, which communicates with the instrument’s low-level data acquisition software. An earlier version of COM Object, instrument OCX, has been implemented by MaxQuant Real-Time, which has allowed for the identification of peptides and SILAC pairs on-the-fly. In the current paper, the implementation of COM Object, Smart MS² (Spectroglyph LLC, Kennewick, WA), creates a user interface
software that customizes the control of LTQ Orbitrap-based mass spectrometers and expands on the iterative precursor exclusion strategy implemented in AMEx. Several additional algorithms were developed and implemented through Smart MS to improve the efficacy of precursor exclusion: noise filtering based on persistent precursor detection, deconvolution of peptide precursors in multiple charge and isotopic states, an indexed dynamically aligned exclusion list, dynamic exclusion extended for the entire precursor elution profile, a percent chimeric intensity (PCI) filter, and an integrated post-acquisition database search that generates an exclusion list of only identified peptides. The application of the developed algorithms results in deeper probing of the HeLa proteome than possible with the conventional DDA strategy. The algorithms were tested on an LTQ Orbitrap XL as a proof-of-concept, and similar strategies can benefit the performance of more advanced and faster duty cycle mass spectrometry platforms.

2.3 Materials and methods

2.3.1 Materials

HPLC-grade water, acetonitrile, formic acid, and Pierce HeLa tryptic digest standard were acquired from Thermo Fisher Scientific (San Jose, CA, USA). The column was prepared in-house by polymerizing a frit from a 3:1 mixture of Kasil 1 (29.1% potassium silicate solution) from PQ Corporation (Valley Forge, PA, USA) and formamide from Sigma-Aldrich (St. Louis, MO, USA) inside a 75 μm internal diameter, 360 μm outer diameter polyimide coated fused silica capillary from Polymicro / Molex (Phoenix, AZ, USA). The frit was cut to approximately 0.2 mm, and the end was polished. The fritted capillary was packed with Magic C18AQ, 3 μm diameter, 200 Å pore size beads from Michrom Bioresources (Auburn, CA, USA). Liquid chromatography (LC) separation was performed on an Ultimate 3500 system from Thermo Fisher Scientific (Sunnyvale, CA, USA), and mass spectrometry data was acquired on an LTQ Orbitrap XL ETD (Tune Plus version 2.5.5) from Thermo Fisher Scientific. The sample was electrosprayed using a distal coated 20 μm internal diameter, 360 μm outer diameter tip with a 10 μm opening from New Objective (Woburn, MA, USA) connected directly to the column head using a Teflon sleeve.
2.3.2 Liquid chromatography

Reversed-phase liquid chromatography was carried out using mobile phase A: 0.1% formic acid in water and mobile phase B: 0.1% formic acid in acetonitrile. 20 μg of HeLa digest were dissolved in 100 μL of 2% acetonitrile, 0.1% formic acid in water. 1 μL (200 ng HeLa lysate) was injected directly onto the analytical column using 2% B at 250 nL/min. The sample was loaded and desalted over 10 minutes in 2% B and separated at 250 nL/min flow rate using the following linear gradient: 2-37% B over 120 minutes, 37-95% B over 10 minutes, 95% B hold for 9 minutes, 95% to 2% B in 1 minute, 2% B hold for 10 minutes.

2.3.3 Mass spectrometry

Data were acquired on an LTQ Orbitrap XL mass spectrometer using matching parameters for conventional analysis enabled through Xcalibur (v. 2.0.7) and intelligent, user-defined DDA analysis enabled through Smart MS² (Spectroglyph, LLC. Kennewick, WA). Full precursor scans were acquired over a 400-1700 Th range in the Orbitrap at 60k resolution (at 400 Th), with the AGC set to 1,000,000 and 500 ms maximum ion accumulation time. The 10 highest intensity eligible precursors (Top 10) were serially selected in 2 Th wide isolation windows centered on the monoisotopic peak and activated for a maximum of 30 ms by CID at 35% normalized energy. In fragment ion accumulation, the AGC target was set to 40,000, the maximum injection time was 100 ms, and the ion trap was scanned at 0.5 Da resolution. Noise filtering and exclusion of previously (during the same LC-MS run) fragmented precursors were accomplished differently between the two approaches. In the conventional analysis, ions with a +1 charge state and lower than 500 intensity units were considered chemical noise, and precursors were excluded for 60 seconds after fragmentation to prevent redundant sampling. In Smart MS² analysis, precursors with a charge of +2 or higher and at least 2 isotopic peaks detected in at least 3 consecutive precursor scans were considered fragmentation candidates (not noise). In Smart MS² analysis, previously fragmented precursors in all isotope and charge states were excluded for a minimum of 30 seconds with the exclusion period extended until the end of the precursor’s elution profile.
2.3.4 Exclusion of identified peptides

Completion of data acquisition for each LC-MS run triggered execution of an in-house developed Python script based on the Pyteomics library\textsuperscript{26} to perform an X!Tandem\textsuperscript{27} (version 2013.09.01.1 Sledgehammer) search against the UniProt human database described below, appended with an equal number of reverse decoy sequences. Carbamidomethylation of cysteine was set as a static modification and up to one tryptic missed cleavage was allowed. The error tolerances were set at 10 ppm for the precursor mass (with the monoisotopic peak mismatch enabled) and 0.4 Da fragment mass error tolerance. The scored spectral matches were filtered to 1% FDR, and the accepted precursors remained in the Smart MS\textsuperscript{2} exclusion database. The precursors which were not identified were removed from the exclusion list, which made them eligible for re-fragmentation in subsequent runs. Smart MS\textsuperscript{2} data acquisition was initiated 8 minutes after injection and terminated at 140 minutes (before column washing and equilibrating) to allow sufficient time to complete the database search and exclusion list generation before the subsequent replicate.

2.3.5 Data analysis

The acquired data was analyzed in Proteome Discoverer 1.4 (Thermo Fisher Scientific) with Sequest HT\textsuperscript{28} peptide spectral match scoring and Percolator\textsuperscript{29} validation and filtering (q < 0.01), and with MaxQuant (v. 1.5.2.8) with 1% FDR filtering.\textsuperscript{30} Both searches were conducted against a human UniProt database containing canonical proteins and known variants from March 2014 appended with 47 common contaminants (88,894 total entries). The maximum precursor mass error was set to 10 ppm and the maximum fragment mass error was set to 0.6 Da. Carbamidomethylation of cysteine was set as a static modification, and oxidation of methionine and deamidation of glutamine and asparagine were set as dynamic modifications. Up to 2 missed tryptic cleavages were allowed. Evaluation of overlapping identifications and cumulative identifications were carried out in Excel.
2.4 Results and Discussion

2.4.1 Precursor selection algorithms applied through Smart MS\textsuperscript{2}

The Smart MS\textsuperscript{2} flow-control process is diagrammed in Figure 2.1. Elution profiles of all fragmentation candidates were monitored in real-time by Smart MS\textsuperscript{2} software. Once fragmented, precursors in all charge and isotopic states were excluded for a user-defined minimum exclusion period (30 seconds in the current experiment). If the software determined that an elution profile of a previously fragmented precursor ion had extended past the minimum exclusion period, the exclusion time of that precursor ion was then automatically extended until elution of the precursor was complete. Non-redundant fragmentation of precursors between replicate experiments was accomplished by an indexed exclusion list where retention times of precursor ions were dynamically aligned during data acquisition, based on the retention times of automatically selected anchor peptides using a 2 minute retention time adjustment window. Anchor peptides were automatically identified by Smart MS\textsuperscript{2} from the initial experiment in each series by selecting the highest intensity ion species in each 5-minute long interval and identifying their chromatographic peak apex positions. The percent chimeric intensity filter (PCI)\textsuperscript{7} was set to 500%, meaning that if a second precursor appeared within the isolation window of an excluded precursor, this second precursor would be considered an eligible candidate for MS\textsuperscript{2} only if its intensity is at least 5 times greater than the intensity of the excluded precursor. This threshold was selected based on the definition of chimeric spectra used in S. Houle \textit{et al.}, which stated that the spectrum match score is significantly reduced at a PCI ≥ 20%. Since we are interested in the “contaminant” precursor, the inverse value is used.\textsuperscript{7}
The flow control diagram of the Smart MS\textsuperscript{2} algorithm (left side) and programmatic interfacing to Thermo's COM Object (right side). The latter incorporates interface "ILTQInstCtrl", which controls the instrument and dispatch interface "_ILTQInstCtrlEvents" used to receive instrument status updates. Upon receiving an event from "_ILTQInstCtrlEvents", Smart MS\textsuperscript{2} adds a list of m/z and intensity data from the latest scan to the previously generated dynamic list, performs real-time isotope deconvolution, and then updates the dynamic list of elution profiles, which also incorporates information on the retention time and charge states of the deconvoluted features. A feature represented by, at least, C\textsubscript{12} and C\textsubscript{13} isotopes is considered to be an elution profile (non-noise precursor) if observed in, at least, two consecutive MS\textsuperscript{1} scans at a mass accuracy of 10 ppm. The algorithm performs real-time retention domain alignment using higher-intensity anchor peptides, and then determines whether an elution profile is off the exclusion list, and the profile intensity matches user-defined criteria (e.g., Top 10, Bottom 10, Middle 10-20, etc.) for collisional activation. If the above conditions are met, the profile of interest is added to the MS\textsuperscript{2} attention list to be passed over the instrument through the "ILTQInstCtrl" interface. Redundant elution profiles of the same species represented by lower intensity charge states are excluded. Upon completion of the MS\textsuperscript{2} event, the profile is added to the dynamic exclusion list (which can be expanded with profiles from previous experiments) and then monitored real-time during the experiment. Once eluted off the column, the profile is dynamically removed from the exclusion list. If an elution profile, which is not on the exclusion list, overlaps with nearly isobaric excluded profile in the retention time domain and conforms to percent chimeric intensity (PCI) filter and user-defined criteria for precursor ion fragmentation (e.g., Top 10, etc.), the elution profile would be chosen for MS\textsuperscript{2} fragmentations. In summary, Smart MS\textsuperscript{2} algorithm monitors all the elution profiles concurrently, dynamically puts them on or removes from the exclusion list, performs real-time alignment of retention time domains and conducts MS\textsuperscript{2} experiments based on user-defined activation criteria.
**Figure 2.2A** presents an example of a precursor being recovered despite co-elution with a near-isobaric (*i.e.* within the same 2 Th isolation window) previously fragmented precursor using the PCI filter. Conventional dynamic exclusion could mask the presence of near-isobaric co-eluting species by acquiring one spectrum when one species may be underrepresented and then excluding both precursors until both have eluted. Compounding dynamic exclusion with the exclusion of precursors from previous experiments can quickly deplete the allowed sampling space (a situation where most m/z windows are ineligible for precursor ion selection for most of the LC-gradient to prevent redundant fragmentation of previously selected precursors). The implemented PCI filter recovered sampling of 300-900 precursors in each run. While this translates to less than 10% of the total acquired MS² scans, without the PCI filter the number of acquired MS² spectra is significantly decreased at the third iteration (data not shown). This decrease is delayed until the fourth iteration by application of the PCI filter.

The Smart MS² noise filtering strategy permits selection of very low abundance features while the conventional intensity threshold approach does not consider features below the intensity threshold. The filter implemented in Smart MS² relies on the persistent appearance of the isotopic envelope of the precursor in at least three sequential precursor scans independent of intensity. After detection in 2 consecutive scans, the precursor is considered a candidate and may be selected for fragmentation in the third scan if it is present. At least 2 isotopic peaks (*i.e.* ¹²C and ¹³C peaks) were considered an isotopic envelope, because, for many low-abundance features, the M+2 peak (*i.e.* 2 x ¹³C) was not consistently detectable. This approach allowed the mass spectrometer to sample precursors which would be considered noise by a conventional threshold filter and, in some cases, acquire informative MS² spectra.

Smart MS² real-time precursor ion charge deconvolution treats all m/z peaks derived from the same peptide species as one entity, so that higher intensity ion species, *e.g.* [M+2H]²⁺, would be fragmented, while lower intensity species, (*e.g.* [M+3H]³⁺ and [M+4H]⁴⁺) would be excluded. In addition, extended dynamic exclusion prevents redundant fragmentation of a precursor for a minimum exclusion duration that is reset if the precursor is persistently detected after the minimum set duration. This is in contrast to conventional DDA where a set exclusion duration time, which is the same for every precursor, is the only means to control redundant fragmentation of precursor ions.
The most important feature of Smart MS\textsuperscript{2} is the automatic alignment of the indexed exclusion list during data acquisition using a set of high-intensity anchor peptides automatically selected from the first run. The database of precursors to be excluded is dynamically adjusted by the difference between the reference and observed retention times of the anchor peptide chromatographic apexes. Figure 2.2B demonstrates the agreement in the on-the-fly alignment of the exclusion list performed by Smart MS\textsuperscript{2} with the pose-cluster alignment\textsuperscript{31} implemented in OpenMS\textsuperscript{32} of the reference LC-MS run (the first iteration of exclusion series) and each subsequent LC-MS iteration in the series. While the two alignment strategies are inherently different, the general agreement indicates that the adjustments made by Smart MS\textsuperscript{2} adequately address the retention time shifts between replicate experiments.

The post-acquisition database analysis, triggered by Smart MS\textsuperscript{2} at the end of each LC-MS experiment, executed an X!Tandem database search. The peptides identified at FDR < 1% remained in the exclusion list, while the unidentified precursors were removed from the exclusion list and were made eligible for fragmentation in subsequent runs. A relatively short time interval during column washing and equilibration was allocated for the database search so X!Tandem was used because it is a fast and readily available open access search algorithm. To further increase the benefits of the post-acquisition database search and exclusion of only identified peptides more comprehensive search strategies can be implemented in the future.
Figure 2.2 – PCI filter (A) and on-the-fly exclusion list alignment (B).
Panel A is an example of a candidate (YQAVTATLEEK, orange) which would not be eligible for fragmentation due to co-elution with a candidate that has been previously fragmented (ALIGYADNQCK, blue) within the same 2 Th isolation window. However, the precursor is fragmented when its intensity surpasses the previously fragmented candidate by the 500% PCI threshold. Panel B demonstrates the agreement between Smart MS² on-the-fly retention time adjustment of the indexed exclusion list and pose-clustering alignment relative to the first run in the exclusion series. Pose-clustering did not align the earliest and latest regions of the gradient due to high variability in elution of the most hydrophilic and hydrophobic peptides.

2.4.2 Performance metrics of Smart MS² and conventional DDA

Three DDA strategies were compared: (1) conventional Xcalibur driven Top 10 approach, to be referred as **Strategy 1**, (2) Smart MS²-driven analysis in which all previously fragmented precursors were iteratively excluded in subsequent replicate LC-MS/MS analyses, to be referred as **Strategy 2**, and (3) Smart MS²-driven analysis, in which only identified peptides (FDR < 1%) were iteratively excluded in subsequent replicates, to be referred as **Strategy 3**. The differences in precursor selection algorithms implemented in the three data acquisition strategies are summarized in **Table 2.1**. Strategy 1 was evaluated using 8 LC-MS/MS runs to approximate the saturation of analysis (the maximum number of
peptide species identified using the analytical platform) in duplicate using two batches of HeLa lysate (Replicate 1 and Replicate 2). Three series of four iterations were used to evaluate the performance of Strategy 2. In a Strategy 2 series the precursors fragmented during the first iteration (i.e. the initial analysis of the sample) were excluded from selection in the second iteration (i.e. second replicate LC-MS/MS analysis of the same sample), the precursors fragmented in the first and second iteration were excluded in the third iteration, and all previously fragmented precursors were excluded in the fourth iteration. Three Strategy 2 series were benchmarked against the extrapolated Strategy 1 saturation. Finally, Strategies 2 and 3 were compared using the same four-iteration series experimental design, in triplicate.

<table>
<thead>
<tr>
<th>Precursor selection logic</th>
<th>Strategy 1: Conventional DDA</th>
<th>Strategy 2: Smart MS$^2$ iterative exclusion of all precursors</th>
<th>Strategy 3: Smart MS$^2$ iterative exclusion of identified peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity based precursor selection</td>
<td>10 highest intensity eligible precursors from each MS$^1$ scan are selected for MS$^2$ (Top 10)</td>
<td>Charge and defined elution profile (isotopic envelope and consecutive scan appearance)</td>
<td>Precursor becomes an eligible MS$^2$ candidate if PCI filter threshold is exceeded</td>
</tr>
<tr>
<td>Noise filtration</td>
<td>Charge and intensity threshold</td>
<td>N/A</td>
<td>60 sec. exclusion after MS$^2$</td>
</tr>
<tr>
<td>Percent chimeric intensity (PCI) filter</td>
<td>N/A</td>
<td>Precursor becomes an eligible MS$^2$ candidate if PCI filter threshold is exceeded</td>
<td>30 sec. minimum m/z exclusion extended until end of elution</td>
</tr>
<tr>
<td>Intra-experiment (dynamic) exclusion</td>
<td>60 sec. exclusion after MS$^2$</td>
<td>All precursors fragmented in previous experiments</td>
<td>Only peptides identified in previous experiments</td>
</tr>
<tr>
<td>Inter-experiment exclusion</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Post-acquisition database search</td>
<td>N/A</td>
<td>N/A</td>
<td>X!Tandem (1% FDR)</td>
</tr>
</tbody>
</table>

Table 2.1 – Precursor selection logic implemented in each data acquisition strategy.

Table 2.2. presents the performance metrics of the tested data acquisition strategies. It is possible to schedule exclusion of specified m/z ranges in Xcalibur, but our attempt to implement exclusion of all identified peptides between two replicate analyses of complex proteomic samples found no benefit in this strategy compared to simply performing two replicate analyses (Figure 2.6). To be effective, exclusion of previously fragmented or identified precursors requires exclusion of all charge and isotopic states and highly reproducible chromatography (or sufficiently wide retention time exclusion windows) as with AMEx, or on-the-fly exclusion list alignment as in our approach. Thus, exclusion was not implemented in the two sets of experiments evaluating the saturation of Strategy 1 (conventional DDA). The identification results of Strategy 1 improved in the second replicate, as indicated by an increase in MS$^2$ scans, peptide
spectral matches, and peptide identifications, which can possibly be attributed to differences in sample batches or experimental variables. However, even with the improved performance, the cumulative peptide identification increased by less than 4% over 8 runs. Smart MS\textsuperscript{2} performance in the first exclusion iteration (no inter-experiment precursor exclusion is applied) acquired fewer precursor (MS\textsuperscript{1}) and MS\textsuperscript{2} scans than the conventional analysis, due to (i) the inability to perform parallel MS\textsuperscript{1} and MS\textsuperscript{2} experiments with Thermo’s COM Object, and (ii) latency in data transfer between the external and internal PCs controlling the mass spectrometer. Evaluation of MaxQuant Real-Time has also observed these latency issues.\textsuperscript{25} However, the ratio of acquired MS\textsuperscript{2} scans to precursor scans was higher in the Smart MS\textsuperscript{2} experiments: 6.3, 7.0, and 7.2 (Strategy 2 replicates 1 and 3 and Strategy 3, respectively) compared to 4.7 and 3.4 in Strategy 1 (see Table 2.2). This suggests that the Smart MS\textsuperscript{2} data acquisition approach was more efficient while somewhat slower in acquiring MS\textsuperscript{2} scans.

2.4.3 Effects of iterative exclusion (Strategy 2) on precursor selection

Iterative exclusion implemented in Smart MS\textsuperscript{2} effectively prevented redundant fragmentation of precursors between technical replicates. Figure 2.3A demonstrates the overlap in peptides identified by Sequest HT with Percolator rescoring and filtering (q < 0.01, equivalent to FDR < 1%) between 3 runs acquired using Strategy 1 (left) and Strategy 2 (right). The typical Strategy 1 overlap in peptide identifications between runs of roughly 70% is decreased to 10-20% when iterative exclusion is applied (Table 2.3). Iterative exclusion also affects the intensity distribution of features sampled. Figure 2.3B was generated through label-free quantitative analysis (MaxQuant), which (1) identified all peptide-like LC-MS features (at least 3 peaks in the isotopic envelope, and charge +2 to +5), (2) determined which features were fragmented, (3) identified peptides using the Andromeda search engine, and (4) filtered to FDR < 1%. The peptide-like features were sorted by intensity into twenty 5-percentile bins, and the percentage of sampled and identified peptide-like features in each bin was determined. The orange line represents the combined feature identifications from all 8 Strategy 1 runs with the expected high-intensity bias observed for conventional DDA in other studies.\textsuperscript{8} The Strategy 2 iterations, presented as an average of 3 experiments, show a reduction in sampling bias with each iteration of precursor exclusion. The first iteration resembles the Strategy 1 distribution; the second iteration shows a shift toward sampling of
medium intensity features, and the third and fourth iterations show evenly distributed sampling across the intensity range. As expected, the ratio of identified to sampled features correlated with precursor ion intensity in both Strategies 1 and 2 (not shown), because low-intensity precursor ion signals are less likely to produce an identifiable fragmentation spectrum. Another caveat of iterative exclusion is the eventual depletion of the eligible MS² candidate pool, as manifested in a decrease in the number of acquired MS² scans and identified unique peptides in the third and fourth iterations (Table 2.2).
Table 2.2 – Performance metrics for each data acquisition strategy.
The performance of Strategy 1 was evaluated using two batches of HeLa lysate (Replicate 1 and 2) and the results are presented as an average of 8 runs and cumulative results from 4 and 8 runs. Strategy 2 was evaluated using the Replicate 1 HeLa lysate batch and the results are presented as an average of 3 series. Strategies 2 and 3 were compared using the Replicate 3 batch of HeLa lysate and the results are again presented as an average of 3 series.
Table 2.3 – Overlap in peptide identifications between replicate LC-MS experiments.
Strategy 1 Replicates 1 and 2 demonstrate typical overlap determined from the first four replicates in each set. The Strategy 2 and 3 results are averaged across 3 series. The low overlap in the second iteration of Strategy 2 indicates that almost 90% of the peptides identified in the second iteration were not detected in the first iteration. In Strategies 2 and 3 the overlap increases in subsequent iterations because the population of previously observed peptides grows while the number of identified peptides decreases due to increased sampling of low abundance species.

<table>
<thead>
<tr>
<th>Experimental Condition</th>
<th>2nd Iteration, Peptide IDs</th>
<th>3rd Iteration, Peptide IDs</th>
<th>4th Iteration, Peptide IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overlap</td>
<td>Cumulative</td>
<td>Overlap</td>
</tr>
<tr>
<td>Strategy 1 (Rep. 1)</td>
<td>79.4%</td>
<td>8,322</td>
<td>88.4%</td>
</tr>
<tr>
<td>Strategy 1 (Rep. 2)</td>
<td>79.2%</td>
<td>8,395</td>
<td>88.2%</td>
</tr>
<tr>
<td>Strategy 2 (Rep. 1)</td>
<td>12.5% +/- 1.4%</td>
<td>8,666 +/- 290</td>
<td>30.7% +/- 10.0%</td>
</tr>
<tr>
<td>Strategy 2 (Rep. 3)</td>
<td>11.0% +/- 1.4%</td>
<td>8,281 +/- 186</td>
<td>25.2% +/- 4.5%</td>
</tr>
<tr>
<td>Strategy 3 (Rep. 3)</td>
<td>33.6% +/- 1.4%</td>
<td>8,539 +/- 484</td>
<td>55.8% +/- 4.6%</td>
</tr>
</tbody>
</table>
Figure 2.3 – Effects of iterative exclusion of precursor selection.
In conventional analysis (Strategy 1), over 70% of identified peptides are redundant between replicates (Panel A, left). In iterative exclusion (Strategy 2) this overlap is reduced to 10-30% (Panel A, right). This decrease in redundant fragmentation allows the MS instrument to target medium and low abundance peptides as shown in Panel B, where the blue line presents the sampling coverage by Strategy 1 from 8 replicate runs and the bar graphs represent the iterations of Strategy 2 (average of 3 runs).
Figure 2.4 – Comparison of identifications: Strategy 1 vs. Strategy 2.
Panel A presents the cumulative peptide identifications using either Strategy 1 or Strategy 2. Each 4-iteration Strategy 2 series identified more peptides than four replicates of Strategy 1. The logarithmic extrapolation of Strategy 1 from 8 runs estimates the saturation at approximately 12,500 unique peptides (light blue line). Combined results from two Strategy 2 series exceed Strategy 1 saturation by ~1,500 unique peptides. Combined results from all 3 Strategy 2 series show that peptide identification does not plateau after over 16,000 peptides identification due to sampling and identification of low-intensity peptides. Panel B presents the overlap in protein group identifications between Strategy 2 series, and the combined identifications between Strategies 1 and 2 with single peptide identifications accepted on the left and a 2 peptide minimum filter applied on the right.

2.4.4 Differences in identifications between Strategies 1 and 2

Figure 2.4A presents the cumulative peptide identifications using Strategy 1 (Replicates 1 (dark blue) and 2 (purple)) and Strategy 2. Extrapolation of experimental data (dotted lines) shows that Strategy 1 would reach the saturation level of approximately 12,500 peptides (dotted light-blue line) after 12+ replicates. After four iterations of Strategy 2, 84.4 +/- 2.2% of all peptides identified at the approximate saturation level are identified. When peptide identifications from two Strategy 2 series are combined, the
saturation level is exceeded by 13%. When results from three Strategy 2 series are combined, the saturation is exceeded by 29% (16,156 total peptide identifications). There is a noticeable decrease in the gain of new identifications in the fourth iteration of Strategy 2 in comparison to preceding iterations. While the fourth iteration provides a substantial number of novel peptide identifications, if the identifications from only the first three iterations in each Strategy 2 series are combined then a total of 15,168 peptides (21% above saturation) are identified over 9 runs (not shown), providing a quicker alternative for in-depth analysis. **Figure 2.4B** presents the overlap between the cumulative protein group identifications using Sequest HT and Percolator (q < 0.01, equivalent to FDR < 1%) in each of the Strategy 2 series (top left Venn diagram) and the overlap in protein identifications between Strategy 1 and Strategy 2 (bottom left Venn diagram). The same Venn diagrams are presented in the right panel when a minimum of 2 peptides per protein group filter is applied in addition to q < 0.01 for more stringent protein identifications. Without the 2-peptide filter, over 1,000 more protein groups are identified with Strategy 2. However, this advantage diminishes to several hundred when the filter is included. Many single-peptide protein groups were identified with high confidence in multiple Strategy 2 series as reflected by the drop in overlapped protein group identifications when the 2 peptide filter is applied, which improves confidence in the identifications despite reliance on a single peptide. The same data was also processed using MaxQuant (FDR < 1%). MaxQuant identified a total of 2,963 protein groups with 304 protein groups unique to Strategy 2 and 23 protein groups unique to Strategy 1. The majority of proteins unique to either strategy are within the lowest 20 intensity percentile (data not shown). Depending on the stringency of data analysis (minimum of either 2 or 1 peptide per protein group, q < 0.01), Strategy 2 identified between 300 to 1000 additional low abundance protein groups undetected by Strategy 1, indicating that the dynamic range of analysis is increased with more efficient data acquisition. The low abundance protein identifications vary between the Strategy 2 series suggesting that the ability to detect these low abundance species is stochastic even with efficient precursor sampling, so the flexibility of an untargeted approach (i.e. Strategies 2 and 3) may be better for discovering these species than the targeted strategies described in the introduction (e.g. PAnDA).
Figure 2.5 – Comparison of identifications: Strategy 3 vs. Strategy 2.
Exclusion of identified precursors (Strategy 3) identified on average (n=3) ~1,000 (or 10%) more peptides over 4 iterations than the exclusion of all fragmented candidates (Strategy 2) (Panel A). The identification gains were hampered by limitations of the post-acquisition search resulting in higher redundant peptide identifications between replicates. The majority of peptides redundantly identified between the first and second iteration of Strategy 3 were not identified in the post-acquisition (X!Tandem) search (Panel B).

2.4.5 Iterative exclusion of identified precursors vs. all fragmented precursors

One of the major deficiencies of Strategy 2, presented in Table 2.2, is a drop in MS² scan acquisition in the third and fourth iterations caused by the depletion of the candidate pool. Many ion precursors already fragmented in the initial LC-MS iteration(s) which were therefore excluded from fragmentation in subsequent iterations did not result in informative MS² spectra and hence, successful peptide identifications. In Strategy 3, an automated database search is executed at the completion of each LC-MS using an in-house developed Python script to exclude only precursors identified by X!Tandem (FDR < 1%) is implemented to address this issue. Strategy 3 shows uniform numbers of MS² scans acquired across all four LC-MS iterations and allows a second opportunity to acquire better quality spectra for previously fragmented, but not identified precursors. Presented in Figure 2.5A, is a significant gain (p < 0.05) of 1,000 peptides (~10%) over four exclusion iterations in Strategy 3 compared to Strategy 2. The performance of Strategy 3 can be improved further with a more comprehensive automated search. As presented in Table 2.3 the overlap in identified peptides between the first and second iterations is over 30% in Strategy 3 (compared to ~10% with Strategy 2). This is due to redundant fragmentation of
peptides that were not identified by the post-acquisition X!Tandem search but are identified by Sequest HT and Percolator (Figure 2.5B). While a more thorough post-acquisition database search would improve performance by reducing the overlap between iterations to the ~10% level observed with Strategy 2, we were limited by the system resources of the instrument controlling PC and the allotted search time (duration of column cleaning and equilibration). Thus, X!Tandem was used without the benefits of post-search validation and re-scoring with Percolator. However, even under suboptimal conditions Strategy 3 is superior to Strategy 2 and demonstrates even greater potential after addressing the above-mentioned limitations.

2.5 Conclusion

This work has demonstrated that improved precursor ion selection increases the number of identified peptides in DDA analysis, which results in the identification of low abundance proteins, previously undetectable with the conventional approach (Top N DDA, Strategy 1). The conventional strategy reaches saturation due to an inefficient sampling of medium and low abundance species in proteome-complexity samples, while the application of efficient data acquisition provides deeper proteomic profiling. Specifically, to increase the efficiency of iterative DDA analyses, we have developed combined real-time tracking and intelligent handling of all elution profiles of isotopically and charge-state deconvolved precursor ions signals. The proposed strategy includes a dynamic (signal-to-noise and persistent precursor detection based) exclusion list of elution profiles of all precursor ion candidates, alignment of retention time domains of different LC runs, non-redundant triggering of MS² events using the most abundant charge states of the peptides of interest, use of a percent chimeric intensity (PCI) filter, and automated post-data acquisition database searching performed immediately after completion of the run (integrated into the Smart MS² workflow to exclude identified precursors from the following analyses). Without the post-acquisition database search, 29% (16,156 vs. ~12,500, FDR < 1%) more peptides were identified above the saturation level of conventional data acquisition after 12 runs. The post-acquisition search improved the performance of iterative exclusion by an average gain of 1,000 peptides (~10%) over 4 exclusion iterations. While a direct comparison is beyond the scope of this publication, these
improvements in identifications are comparable to those reported for AMEx and PAnDA, (26% and 31% peptide identification gains respectively over 6 iterations).

Recent generations of mass spectrometers are certainly capable of faster precursor sampling rates than an LTQ Orbitrap XL. However, these sampling rates are still insufficient for fragmentation of every eluting precursor in complex samples. Furthermore, newer mass spectrometers also tend to offer higher sensitivity, which translates to more detectable candidates that require even faster sampling rates for comprehensive sampling. As such, even the most advanced mass spectrometers would benefit from improved precursor ion selection strategies such as those presented in this proof-of-concept study. Alternatively, DIA allows comprehensive sampling, but identification of low abundance precursors from multiplexed spectra can be problematic. While numerous software packages for interpreting DIA data are now available, the fundamental obstacle of accumulating sufficient signal for fragments from low-intensity precursors co-fragmented with high-intensity precursors is still a limitation. Comparisons of DDA and DIA analysis demonstrated the two approaches identify complementary peptide populations. Furthermore, DIA data analysis by targeted data extraction (peptide-centric analysis) requires a comprehensive spectral library that is typically obtained through exhaustive DDA analysis of the sample of interest. The described approach can obtain spectra for peptides beyond the saturation point of conventional DDA in LC-MS proteomic profiling with replicate injections, allowing the opportunity for detection of these low abundance species in DIA analysis. The depth of the proteomic coverage is commonly increased by higher efficiency separation of analytes using multi-dimensional and high-resolution liquid chromatography. The efficient exclusion of previously fragmented precursors allows an additional dimension for pseudo-separation by manipulating the precursor ion sampling in MS data acquisition. These developed precursor ion selection and exclusion algorithms will be instrumental in any experiments where liquid phase separation (e.g. liquid chromatography, capillary electrophoresis, capillary isoelectrofocusing, etc.) is coupled to mass spectrometry to enable deep molecular profiling (i.e. proteomic, lipidomic, metabolomics, etc.) of complex biological samples and, especially limited samples, where all steps towards increasing the coverage of profiling are important.
The performance of the developed algorithms was noticeably decreased by latency in data transfer between the data acquisition PC and the on-board computer. Additionally, the post-acquisition database search was limited by the PC system resources. We expect to see even greater advantages in sample analysis by the developed precursor ion selection algorithms with improvements in computational hardware. In summary, better precursor selection can provide a gain in depth of proteomic profiling by reducing the inefficiencies of conventional precursor sampling. Similar gains can be expected in the analysis of post-translational modifications when modified peptides are adequately enriched. It should be noted that quantitation by spectral counting is not possible when high abundance precursor bias is reduced by iterative exclusion. However, we anticipate that quantitation by precursor peak integration will be further improved by identification of additional peptides per protein. The DDA algorithms reported here were implemented in the Smart MS² platform, which was developed in collaboration with, and can be licenced from Spectroglyph LLC. The required LTQ COM Object instrument control library can be licensed from Thermo Fisher Scientific. The Python script enabling post-acquisition database search is available upon request from the Barnett Institute.

2.6 Assessment of Xcalibur driven exclusion

Precursor exclusion implemented through the conventional instrument control (Xcalibur) was tested by examining the overlap in peptide identifications between an initial run, a second run, and a run where all peptides identified in the initial run (Sequest HT with Percolator q < 0.01) are excluded. As shown in Figure 2.6, the experiment found that the exclusion implemented through the standard control did not efficiently exclude identified peptides, and there was no benefit in cumulative identifications, which supported our decision not to use the exclusion feature when evaluating conventional DDA (Strategy 1).
Figure 2.6 – Xcalibur driven exclusion does not improve peptide identifications.
The overlap in identified peptides between technical replicate 1 and replicate 3, which excludes peptides identified in replicate 1 is lower than the overlap between replicate 1 and replicate 2 (Panel A). However, the relatively low number of total identified peptides does not result in additional peptide identifications compared to not using exclusion (Panel B).

2.7 Acknowledgements

This work was supported by the Barnett Institute of the Chemical and Biological Analysis at Northeastern University, ASMS 2015 Research Award (ARI), and the Dana-Farber Cancer Institute and Northeastern University Joint Seed Funding Program in Cancer Drug Development (ARI). We thank Thermo Fisher Scientific for providing access and licensing the LTQ COM Object instrument control library. This is contribution number 1058 from the Barnett Institute.
2.8 Reference


8. Michalski, A.; Cox, J.; Mann, M., More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *Journal of Proteome Research* 2011, 10, (4), 1785-1793.


Chapter 3: Comprehensive HCP Profiling by Targeted and Untargeted Analysis of DIA Mass Spectrometry Data with PRM Verification

This chapter is a formatted version of an article, with the same title, submitted to Analytical Chemistry. The author of the dissertation wrote the manuscript, developed the implemented workflow, conducted a portion of the experimental work, and processed the data. Dr. Yuanwei "Abby" Gao assisted with sample preparation, assembly of the targeted assay library, and manuscript preparation. Somak Ray automated the data processing procedures. The samples were generated at Bristol-Myers Squibb by Dr. Mi Jin, Dr. Zhijun Tan, Dr. Nesredin A. Mussa, Dr. Li Tao, and Dr. Zhengjian Li. Dr. Alexander Ivanov provided manuscript feedback. The entire project was supervised by Professor Barry Karger, who also contributed significantly to experimental design and preparation of the manuscript.
3.1 Abstract

Host cell proteins (HCPs) are process-related impurities of biopharmaceuticals that remain at trace levels despite multiple stages of downstream purification. Currently, there is interest in implementing LC-MS in biopharmaceutical HCP profiling alongside conventional ELISA, because individual species can be identified and quantitated. Conventional data-dependent LC-MS is hampered by the low concentration of HCP-derived peptides, which are 5-6 orders of magnitude less abundant than the biopharmaceutical-derived peptides. In this paper, we present a novel data independent acquisition (DIA)-MS workflow to identify HCP peptides using automatically combined targeted and untargeted data processing, followed by verification and quantitation using parallel reaction monitoring (PRM). Untargeted data processing with DIA-Umpire provided a means of identifying HCPs not represented in the assay library during conventional targeted, peptide-centric, data analysis. A monoclonal antibody (mAb) purified by Protein A column elution, cation exchange chromatography, and ultrafiltration was analyzed using the workflow with 1D-LC. Five protein standards added at 0.5 to 100 ppm concentrations were detected in the background of the purified mAb, demonstrating sensitivity to low ppm levels. A calibration curve was constructed based on the summed peak areas of the three highest intensity fragment ions from the highest intensity peptide of each protein standard. 24 HCPs were identified and quantitated based on the calibration curve over the range of low ppm to over 100 ppm in the purified mAb sample. The developed approach achieves comprehensive HCP profiling using a rapid 1D-LC method and specific identification by exploiting the high mass accuracy and resolution of the mass spectrometer.

3.2 Introduction

Host-cell proteins (HCPs) are ubiquitous process-related trace impurities of purified biologically-derived pharmaceutical products.\textsuperscript{1,2} Conventional HCP analysis by ELISA generally uses anti-sera raised against the HCP pool (\textit{i.e.} the host cell proteins expressed by the null cell line).\textsuperscript{3} ELISA analysis provides a bulk quantitation of the overall HCP abundance. However, there is no information on individual species or their concentrations. Currently, there is interest to identify individual species as specific HCPs could be toxic, immunogenic, or potentially degrade the drug substance.\textsuperscript{3,4} Furthermore, specific HCP species may
be immunogenic in humans but may not elicit a strong immune response from the donor animal, leading to an underestimate of the immunogenicity risk by ELISA.\(^5\) In addition, development of a process-specific ELISA method may require months to generate high quality polyclonal antibodies.\(^3\) LC-MS analysis, on the other hand, can identify and quantitate individual protein species at low ppm levels and has the potential to be an orthogonal analytical method to complement or even substitute conventional ELISA.\(^6\)\(^-\)\(^9\)

In comparison to proteomic analysis of high complexity samples, such as cell lysates, HCP analysis involves samples of much lower complexity (fewer than 50 proteins generally, compared to more than \(10^4\) proteins in a cell lysate) with the need for a high dynamic range to detect HCPs at the low ppm level in the presence of the therapeutic protein. Previous LC-MS approaches attempted to address the high dynamic range issue by depleting the therapeutic protein\(^10\), \(^11\) or using multi-dimensional chromatography,\(^12\) with ion-mobility mass spectrometry as an additional separation dimension.\(^13\) Depletion of the therapeutic protein would seem to be a potential solution to the high concentration range challenge; however, extra steps are required in the analysis, and some HCP species may interact with the therapeutic protein\(^14\), \(^15\) and would also be depleted. HCP analysis in the background of the therapeutic protein is, generally, the desired approach both for simplicity and completeness. While some studies have used data-dependent acquisition (DDA) to identify HCPs,\(^16\)\(^-\)\(^18\) data-independent acquisition (DIA) represents a potentially superior alternative.\(^5\), \(^12\), \(^13\), \(^19\), \(^20\) DDA suffers from a bias towards sampling and identification of high abundance species, resulting in poor and inconsistent detection of low level HCPs. The key advantage of DIA is that MS2 fragment information from all eluting precursors is acquired, and identification of proteins, even at low levels, is more reproducible.\(^21\) Contrary to DDA data, DIA data are multiplexed (each spectrum potentially contains fragments from multiple precursors), and fragment ions are detected multiple times across the elution peak, enabling simultaneous chromatographic tracking of the precursor as well as the associated fragment ions.

Targeted (also referred to as peptide-centric) DIA data analysis requires a comprehensive assay library.\(^22\), \(^23\) An assay, in this context, includes the normalized retention time of the peptide, the precursor m/z, and the relative intensities of characteristic MS2 fragment ions (typically \(b\) and \(y\) ions from high energy collision dissociation (HCD) or collision induced dissociation (CID)). Moreover, each peptide
charge state constitutes a separate assay.\textsuperscript{22, 23} Thousands of assays can be generated from a DDA experiment by extracting the most intense fragment ions from high confidence peptide spectral matches. However, in DDA, the number of identifying spectra, and consequently, the number of targeted assays, correlates with protein abundance.\textsuperscript{24} Additionally, many low abundance proteins may not be detected by DDA, resulting in an incomplete library with a bias towards high abundance proteins. The assay library can be enhanced through sample pre-fractionation (2D-LC-MS), additional replicates, analysis of related samples (e.g. null-cell line lysate and samples from early stages of purification), and targeted analysis. Nevertheless, there is always potential that low level HCPs may not be represented in the assay library. Targeted analysis can be supplemented by an “untargeted” strategy in which the DIA data is converted to pseudo-DDA data with DIA-Umpire,\textsuperscript{25, 26} and this data is searched against the entire Chinese Hamster Ovary (CHO) protein sequence database (or other relevant database). A recent benchmark comparison of various DIA data processing strategies found that untargeted analysis with DIA-Umpire identified a significant number of additional peptides beyond targeted analysis in spite of an extensive targeted assay library.\textsuperscript{27} A further challenge of HCP analysis is evaluation of the confidence in the peptide and subsequent protein identifications using standard target-decoy approaches that were developed for high complexity samples. For purified therapeutics, the number of HCP proteins in the sample will generally be fewer than 50, and the true positive peptide population would be too small to accurately determine a false-discovery rate (FDR) threshold, leading to a compromise between identification sensitivity and specificity.\textsuperscript{28}

In this paper, we describe a novel, automatable workflow for rapid in-depth HCP analysis of a monoclonal antibody (mAb) sample after several stages of purification (Protein A column elution, cation exchange chromatography, and ultrafiltration buffer exchange). First, a targeted assay library containing over 4,000 assays for 632 protein groups was generated by 2D-LC-MS DDA analysis of a Protein A column eluate containing the mAb therapeutic. Then, the processed mAb sample, spiked with protein standards spanning a 0.5 - 100 ppm concentration range, was analyzed by LC-DIA-MS using targeted and untargeted data analysis, followed by verification and quantitation by parallel reaction monitoring (PRM). The methodology identified 24 HCPs down to low ppm level in the purified mAb sample, using 1D-LC-MS. The presented data analysis strategy is readily automatable and allows a rapid transition from
DIA to PRM, enabling sensitive and specific identification of HCPs down to the low ppm level. The label-free quantitation based on fragment peak area provides a convenient estimation of HCP abundance suitable for support of downstream process development with potential for more accurate quantitation with stable isotope labeled standards.

3.3 Materials and Methods

3.3.1 Materials and equipment

Samples of a monoclonal therapeutic antibody after Protein A column elution and purified further by cation exchange chromatography and ultrafiltration buffer exchange (purified mAb) were generated from a CHO cell line at the Bristol-Myer Squibb bioprocessing facility (Devens, MA). Triethylammonium bicarbonate buffer (TEAB) (1.0 M, pH 8.0), LC-MS grade ammonium hydroxide solution (25% in H₂O), cytochrome C (≥ 95%, from horse heart), lysozyme (≥ 95%, from chicken egg white), β-casein (≥ 90%, from bovine milk), myoglobin (≥ 95%, from horse skeletal muscle), lactoferrin (≥ 90%, from bovine milk), dithiothreitol (DTT), iodoacetamide (IAM), formic acid, and LC-MS retention time calibration standards, were obtained from MilliporeSigma (St. Louis, MO). The bicinchoninic acid (BCA) protein assay kit, LC-MS grade water, and LC-MS grade acetonitrile were obtained from Thermo Fisher Scientific (Waltham, MA). Sequencing-grade modified trypsin was purchased from Promega (Madison, WI), and MS grade lysyl endopeptidase (Lys-C) from Wako (Richmond, VA). A set of 10 stable isotope labeled (lysine +8 or arginine +10) tryptic peptides (8 - 13 residues) spanning the elution gradient were purchased from JPT GmbH (Germany).

High pH (~10) LC fractionation was carried out off-line using an XBridge peptide BEH C18 column (3.5 µm beads, 300Å, 2.1 x 150 mm) (Waters Corporation, Milford, MA) on an Agilent 1200 LC (Agilent Technologies, Santa Clara, CA). An ACQUITY UPLC M-class peptide CSH C18 column (1.7 µm beads, 130Å, 0.3 x 150 mm) (Waters Corporation) on an Ultimate 3500-RS LC (Thermo Fisher Scientific) was used for low pH (~2.7) reversed phase separation. All MS data were acquired on a QExactive Plus mass spectrometer with a heated ESI source (Thermo Fisher Scientific, San Jose, CA).
3.3.2 Sample preparation

Aliquots containing 0.6 mg of the mAb sample (determined by BCA assay) were denatured in 150 µL of 8 M urea and 100 mM TEAB (pH 8.0). The sample was reduced in 10 mM DTT at 37 °C for 1 hour and then alkylated with 10 mM IAM in the dark and at room temperature for 45 minutes. 900 µL of cold acetone (pre-chilled to -20 °C) were added and the proteins precipitated at -20 °C overnight. After centrifugation at 12,000 x g for 15 minutes, the supernatant was discarded. The precipitated proteins were reconstituted in 250 µL of the digestion buffer, 25 mM TEAB (pH 8.0) in 10% acetonitrile and 90% water, and Lys-C was then added at a 1:100 (w/w) ratio. The samples were digested for 5 hours at 37 °C, and trypsin was added at 1:50 (w/w) to continue digestion overnight (18 hours) at 40 °C. After digestion, the samples were lyophilized to dryness and stored at -80 °C.

3.3.3 LC-MS

The digested purified mAb samples and high pH fractions of the Protein A eluate were injected directly onto an ACQUITY UPLC M-class peptide CSH C18 column and separated at 10 µL/min flow rate. Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. Gradient elution consisted of 10-minute loading and desalting in 2% acetonitrile, followed by a linear ramp to 32% acetonitrile over 120 minutes, then a linear ramp to 90% acetonitrile over 20 minutes, isocratic hold at 90% acetonitrile for 8 minutes, return to 2% acetonitrile in 1 minute, and a 20-minute hold at this mobile phase. The eluent was sprayed at +4.5 kV through the heated ESI source with nitrogen sheath gas. The mass spectrometer precursor scan was acquired at 70,000 resolution (at m/z =200), with the AGC set to 1,000,000 and 110 ms maximum injection time. The 455.12002 m/z polydimethylcyclosiloxane ion was set as the internal lock-mass calibrant. In DDA analysis, the MS1 scan was acquired over 400 – 1600 Th, whereas in DIA and PRM analysis, the MS1 scan was acquired over 390 – 1100 Th. In all data acquisition strategies, MS2 scans were acquired by HCD at a normalized collision energy set to 28, with the AGC set to 50,000 and a maximum ion injection time of 110 ms.

3.3.4 Targeted assay library assembly
Three aliquots of digested Protein A column eluate (0.6 mg each, 1.8 mg total) were combined and reconstituted in 40 µL of aqueous 20 mM ammonium formate (pH 10). The sample was injected directly onto the XBridge peptide BEH column and separated using a water and acetonitrile gradient with a 20 mM ammonium formate (pH 10) modifier at 0.2 mL/min. The gradient was delivered as follows: 2% acetonitrile for 30 minutes for loading and desalting, 46-minute linear ramp to 90% acetonitrile, 2-minute linear drop to 2% acetonitrile, and 9-minute hold at 2% acetonitrile. Fractions were collected every 2.5 minutes after the 30-minute desalting period. The 23 collected fractions were pooled by combining fractions 1, 11 and 21, fractions 2, 12 and 22, fractions 3, 13, and 23, and the remaining fractions were pooled as fractions 4 and 14, 5 and 15, 6 and 16, etc. The 10 resulting fractions were dried in vacuum and reconstituted in 120 µL of 0.1% formic acid (roughly 1.5 µg/µL concentrations), and spiked with 100 fmol of LC-MS retention time calibrant peptides.

Each fraction was injected directly onto the ACQUITY UPLC M-class peptide CSH C18 column, separated using the described LC-MS protocol, and analyzed by Top 15 DDA. The MS2 scans were acquired using 1.5 m/z wide isolation windows centered on the monoisotopic peak at 35,000 resolution (m/z = 200). Each fraction was analyzed in triplicate. In the first replicate, 30 µg of sample were injected onto the column to maximize HCP loading. 15 µg were injected in the second replicate to avoid column over-loading. 15 µg were also injected in the third replicate, but the mass spectrometer was set to exclude precursors that were identified in the second replicate at high confidence (FDR < 0.5%) by an automated Morpheus\textsuperscript{39} post-acquisition search.

The acquired data were processed in SearchGUI\textsuperscript{30} using the Myrimatch\textsuperscript{31} and MS-GF+\textsuperscript{32} algorithms and the CHO-K1 consortium database of 24,044 protein sequences.\textsuperscript{33} Maximum mass errors of 6 ppm were allowed for the precursor and fragment m/z, and carbidomethylation of cysteine was set as a fixed modification. Heavy lysine and arginine (for detection of RT standards) and oxidation of methionine were added as variable modifications. The search allowed for up to 2 missed cleavages and one tryptic miscleavage (semi-specific). The results from both search engines were combined and re-scored by PeptideShaker\textsuperscript{34} and exported in the mzIdentML format. An in-house script was used to extract a maximum of 10 highest intensity b and/or y ion transitions from the top scoring spectrum for each peptide.
in each charge state identified at <1% FDR directly from the mzIdentML file. The script also normalized the retention times in each replicate to the average retention times of the RT standards by linear regression. The output of the script was an OpenSWATH compatible tab separated values (TSV) file containing non-redundant assays from all 30 runs with normalized retention times. The files were translated into the TraML format, and the assay library was appended with an equal population of “shuffled” decoy assays using OpenSWATH\textsuperscript{35} to generate the “targeted assay library” for DIA data analysis. A more detailed description of the data processing can be found in Section 3.6.

3.3.5 DIA to PRM analysis

The purified mAb sample, spiked with 5 protein standards spanning the concentration range of 0.5 – 100 ppm (relative to the therapeutic), was digested and analyzed with 1D-LC-DIA-MS and separately with 2D-high pH/low pH reversed phase LC-DIA-MS using 5 fractions. DIA data were acquired using cycles of a single precursor scan followed by 27 DIA scans acquired at 35,000 resolution (m/z = 200). The DIA scans used 22 Th wide isolation windows covering the 400 – 1,000 Th range, with an 11 Th offset between cycles. The data were analyzed by combined targeted OpenSWATH\textsuperscript{22} and untargeted DIA-Umpire\textsuperscript{25} analysis. The untargeted search converted the DIA data into pseudo-MS2 spectra that were searched against the NCBI CHO protein sequence database\textsuperscript{33} using the same settings as in analysis of DDA data, but with a 5% FDR cut-off. The results of the untargeted search were converted into the untargeted assay library. Both the targeted (from Protein A eluate analysis) and untargeted assay libraries were used to probe the DIA data, and PyProphet\textsuperscript{36} was used to select the best peak (without FDR filtering) for each peptide assay. The putative peptides were filtered to exclude therapeutic mAb, trypsin, Lys-C, and common contaminant derived peptides. The remaining putative HCP and protein standard peptides in all identified charge states were exported as entries into a QExactive “inclusion” list. The automation of this procedure and further details are described in Section 3.6.

The putative peptides were targeted by scheduled PRM analysis using a 1.4 Th wide isolation window with the center offset by +0.4 Th from the monoisotopic m/z and a 10-minute (Experiment 1) or 6-minute (Experiment 2) retention time window centered on the peak apex observed in the DIA analysis.
The instrument acquired one 390 – 1,100 Th MS1 scan for every three PRM scans at 70,000 resolution (at m/z = 200). The PRM data were analyzed in Skyline using a combined assay library containing the targeted assay library and the untargeted search results. A dot product threshold of 0.2 was applied in Skyline to remove the majority of true negatives. All remaining assays were manually examined and were considered true positive matches if at least 4 transitions with mass errors of < 5 ppm were observed co-eluting within ± 3 minutes of the normalized retention time. The fragment peak areas from verified peptides were exported and used to estimate HCP abundance. Peptide signal was calculated as the sum of the 3 highest intensity fragment ions, and the highest intensity peptide was used for protein quantitation. The protein standards were used to construct a linear calibration curve based on concentration using the least squares method. HCP concentration was approximated using the calibration curve, and bulk HCP level was approximated as the sum of individual HCP concentrations multiplied by the average HCP weight estimated at 50 kDa.

3.4 Results and Discussion

3.4.1 Overview of the DIA to PRM workflow for HCP identification and quantitation

We have developed a novel, automatable workflow for HCP identification and quantitation in purified monoclonal antibodies and other biotherapeutics, achieving sensitive detection of HCPs down to the low ppm level using one dimension of LC separation. 24 HCPs were identified and quantitated in a CHO cell generated mAb sample that was purified by Protein A chromatography, cation exchange chromatography, and ultrafiltration. The workflow consisted of four stages as summarized in Figure 3.1: (1) Generation of a high quality assay library for targeted DIA data analysis using high/low pH reversed phase 2D-LC-DDA-MS analysis of the Protein A eluate. (2) DIA analysis of the sample, in this case the purified mAb spiked with 5 protein standards at concentrations of 0.5 to 100 ppm. (3) Identification of putative HCP tryptic peptides in the acquired DIA data using a combination of the targeted and untargeted search strategies. (4) Verification and estimation of the putative HCPs by PRM using tryptic peptides of the spiked-in proteins as internal standards. The methodology takes advantage of the high-resolution capabilities of mass spectrometry instrumentation and is relatively rapid, simple to operate with automated data
processing, and achieves quantitation at the levels necessary for HCP analysis. We next detail the various steps of the workflow and describe the results obtained in analysis of a purified mAb sample.

**Figure 3.1 – Workflow for HCP identification and quantitation.**
Green - LC-MS steps; orange - targeted search steps including assembly of the targeted assay library; purple - untargeted search steps; blue - PRM verification and quantitation steps. Further details about the assembly of the targeted library and the DIA search can be found in Sections 3.3 and 3.6.

### 3.4.2 Stage 1: targeted assay library assembly

In the first stage, as detailed in the Experimental Section, a library containing 4,535 assays corresponding to 3,632 peptide sequences (in different charge states) from 632 proteins, including the mAb, was generated through high/low pH reversed phase 2D-LC-DDA-MS analysis of the Protein A eluate sample using 10 high-pH fractions analyzed in triplicate by low-pH reversed phase LC-MS. The Protein A eluate was sufficiently complex (> 500 protein groups) for the FDR to be accurately set at 1% using the target-decoy model (see below). The retention time was normalized across the 30 runs using spiked-in retention time standards; a requirement since the peptide populations differed between
fractions. Once the retention time normalized assay library was assembled, a specific set of peptides in the library could be used for retention time normalization during DIA analysis of the purified mAb. In our case, 20 high intensity mAb derived peptides spanning the gradient time were selected.

**Figure 3.2** presents the number of assays on a base 2 logarithmic scale incorporated into the targeted assay library for each protein (Y-axis) plotted against the protein intensity determined by MaxQuant\textsuperscript{37} on a base 10 logarithmic scale (X-axis). The figure demonstrates a concentration range spanning more than 6 orders of magnitude between the lowest level HCPs and the mAb. Contrary to an assay library generated from the null cell line, the Protein A eluate library is focused on the HCPs that could potentially be in the further purified samples. However, analysis of the HCPs in the Protein A eluate is affected by the high level of the mAb. As evidence of this, **Figure 3.2** shows a positive correlation between protein abundance and the number of targeted assays. A more complete library could be constructed using more fractions, additional technical replicates, or by supplemental analysis of a null-cell line,\textsuperscript{38} but this would still not guarantee a complete assay library. Untargeted analysis by DIA-Umpire has been shown to identify proteins that are not represented in the assay library in complex samples,\textsuperscript{27} and this principle is exploited in this workflow.
Figure 3.2 – Comparison of protein intensity and number of assays in the targeted assay library. The targeted assay library generated from high/low pH RP 2D-LC-DDA-MS analysis of the Protein A eluate presented as a comparison between the number of assays and protein intensity. The positive correlation suggests that the targeted DIA data processing search alone is inefficient in detecting low level HCPs. The proteins that were detected in the processed mAb sample are circled in red, and the mAb proteins and digestion enzymes are labeled for reference.

3.4.3 Stages 2 and 3: DIA analysis of the purified mAb

With the assay library established, DIA data could be acquired and analyzed in the second and third stages of the workflow (Figure 3.1). As detailed in the Experimental Section, the purified mAb was analyzed by 1D-LC-DIA-MS for rapid analysis, and the acquired data were interrogated by the targeted assay library using, in this case, 20 mAb peptides for retention time normalization (targeted search). In addition, and automatically, an untargeted search was conducted in which the data were converted to pseudo-MS2 spectra using DIA-Umpire, followed by a search against the NCBI CK1 CHO protein sequence database. The data processing procedure was developed to utilize open access algorithms that were weaved together using an in-house script, as detailed in the Supplementary Material. Figure 3.3 presents the number of putative peptide identifications from this combined (targeted and untargeted) analysis of 3 separate runs. Figure 3.3, which will be discussed further below, demonstrates that the two
strategies identify overlapping, as well as different peptide populations supporting the necessity for performing both analyses.

Contrary to high complexity samples, the assessment of false discovery rate (FDR) is not straightforward in a purified sample. Typically a target-decoy model is utilized to evaluate the confidence of the peptide and protein identifications.28, 39 However, a purified mAb is low in complexity (fewer than 50 proteins), and the target-decoy model is problematic because the true positive population (peptides present in the sample) is much smaller than the decoy population consisting of peptides from a scrambled CHO protein sequence database. In targeted DIA data analysis, the FDR is established using a machine learning algorithm such as mProphet to assess attributes of library matches (co-elution, dot product, etc.) and generate a weighted score that best distinguishes between the target and mass shifted or shuffled decoy assays.40 When the true positive population is small, however, the decoy percentage is inflated, and the score threshold cannot be accurately set. If the score is set at the traditional 1% FDR, low scoring but true identifications, which would be accepted with a larger true positive population, will be discarded. Our workflow circumvents this risk by using less stringent FDR filters: 5% FDR for the untargeted search (DIA-Umpire) and only the internal OpenSWATH quality filters without an FDR threshold for the targeted search. This strategy minimized the chance of missing low level HCPs, but consequently, roughly 40% false positive identifications were putatively accepted. The subsequent PRM analysis step was used to verify the true positive peptides and discard false positive identifications.
After generating a list of 154 putative HCP and protein standard peptides from the targeted and untargeted analysis of the 1D-LC-MS DIA data, PRM analysis was employed to verify true positive identifications and filter out false positive matches. The PRM approach isolated precursors in a narrow m/z window (1.4 Th), and the high mass accuracy (5 ppm maximum mass error) and resolution (70,000 at m/z 200) of the Orbitrap for both the precursor and fragment ions were used to identify peptides with high
specificity. The retention time overlap of the precursor and characteristic fragment ions was the basis for accepting a peptide as a true positive. Using Theoretical PRM, we confirmed that a minimum of 3 characteristic fragments at 5 ppm mass error along with a 1.4 Da precursor isolation window could be used to specifically identify a peptide in the CHO protein sequence database. In practice, at least 4 characteristic fragment ions were used to verify peptides in each case.

The number of verified peptides determined from the targeted and untargeted approaches are listed in Figure 3.3 for 3 separate technical replicates. There is a high degree of agreement between the targeted and untargeted search strategies, with 41 - 42 overlapped PRM verified peptides (from a total of 154 putative peptides). At the same time, 16 - 18 peptides were exclusively identified by the targeted strategy and 26 - 28 by the untargeted search, for a total of 85 peptides. This list includes peptides from the spiked-in standards and contaminant proteins. With an expanded targeted library, the number of identifications from the targeted search will likely be increased, but the untargeted search will still, at a minimum, provide a measure of the completeness of the analysis at no additional time or cost. The 16 - 18 peptides observed in the targeted search but not detected by the untargeted approach were likely due to the limitations of the search engines (MSGF+ and Myrimatch) in the low complexity samples. Based on the verified peptides, a total of 24 HCP proteins were identified in the purified mAb sample.

The ability to conduct HCP analysis using 1D-LC has an obvious time advantage compared to 2D-LC approaches. To assess potential losses in sensitivity using only the 1D approach, the purified mAb sample was analyzed by 2D-high/low pH reversed phase LC-DIA-MS with 5 high-pH fractions run in duplicate in the second dimension. A total of 600 mg of digested purified mAb were injected unto the high pH column and an equivalent of 15 µg were injected unto the analytical column with identical settings used for 1D-LC analysis. No additional HCP proteins were found, although several additional peptides were identified for the higher concentration HCPs. The ability to use 1D-LC-MS without significant loss of information is an important advantage for frequent utilization of the workflow.

The 24 verified HCPs are presented in Table 3.1. Experiment 1 was focused on verification of putative HCP peptides. Thus a 10 minute retention time window was used in the PRM analysis to ensure detection of HCP despite potentially large shifts in retention time. Three PRM runs were carried out to
provide a first estimate of concentration of each HCP with %CV for technical replicates. Quantitation is discussed in the next section. Many of the identified HCPs have been reported in previous studies. For example, clusterin,\textsuperscript{12, 19, 42-44} putative phospholipase B-like 2,\textsuperscript{42, 45} 78 kDa glucose-regulated protein,\textsuperscript{12, 19, 42, 44} and protein disulfide-isomerase\textsuperscript{16, 19, 42} have been observed in mAb and Fc fusion protein samples after Protein A purification and CEX chromatography. Other HCPs such as Sister chromatid cohesion protein PDS5, and TRAF3 interacting protein have not been previously reported. Notably, the HCP profile can be affected with slight changes in the mAb structure and upstream and downstream processes.\textsuperscript{14, 46}
While a peptide unique to this protein was identified with high confidence, it is not clear whether the entire protein was present in the sample or a clipped portion.

### 3.4.5 Stage 4B: PRM quantitation

The PRM stage also targeted peptides from the five spiked-in protein standards (β-casein, lysozyme, myoglobin, lactoferrin, and cytochrome C) listed in Table 3.2 along with their concentrations to evaluate the sensitivity and linearity of the workflow and for construction of a calibration curve to estimate the abundance of the identified HCPs. The signal for each peptide was determined as the sum of the 3 highest intensity b or y ion transitions, and the protein concentration was estimated from the highest intensity peptide signal (signals from multiple charge states were combined). Single peptide-based

<table>
<thead>
<tr>
<th>Accession number</th>
<th>Host cell protein</th>
<th>Experiment 1 (n = 3)</th>
<th>Experiment 2 (n = 10)</th>
<th>Verified Peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average (ppm)</td>
<td>%CV</td>
<td>Average (ppm)</td>
</tr>
<tr>
<td>XP_007614348.1</td>
<td>NACHT, LRR and PYD domain-containing protein 14†</td>
<td>130</td>
<td>59%</td>
<td>86</td>
</tr>
<tr>
<td>XP_007636365.1</td>
<td>Putative phospholipase B-like 2</td>
<td>126</td>
<td>12%</td>
<td>89</td>
</tr>
<tr>
<td>XP_007613399.1</td>
<td>Uncharacterized protein C15orf39</td>
<td>87</td>
<td>42%</td>
<td>16</td>
</tr>
<tr>
<td>XP_007616105.1</td>
<td>Long-chain-fatty-acid-CoA ligase 6</td>
<td>39</td>
<td>83%</td>
<td>31</td>
</tr>
<tr>
<td>XP_007619666.1</td>
<td>Lactotransferrin</td>
<td>40</td>
<td>62%</td>
<td>15</td>
</tr>
<tr>
<td>XP_007614428.1</td>
<td>Clusterin</td>
<td>22</td>
<td>2%</td>
<td>32</td>
</tr>
<tr>
<td>XP_007623559.1</td>
<td>26S proteasome non-ATPase regulatory 5</td>
<td>27</td>
<td>45%</td>
<td>34</td>
</tr>
<tr>
<td>XP_007621224.1</td>
<td>TRAF3-interacting protein 1 isoform X2</td>
<td>27</td>
<td>68%</td>
<td>34</td>
</tr>
<tr>
<td>XP_007608950.1</td>
<td>Serine/threonine protein kinase MPCK</td>
<td>24</td>
<td>87%</td>
<td>23</td>
</tr>
<tr>
<td>XP_007639237.1</td>
<td>DNA Repair Protein RAD52</td>
<td>22</td>
<td>13%</td>
<td>13</td>
</tr>
<tr>
<td>XP_003512047.1</td>
<td>Disintegrin metalloprotease 8</td>
<td>20</td>
<td>7%</td>
<td>12</td>
</tr>
<tr>
<td>NP_001233668.1</td>
<td>78 kDa glucose-regulated protein precursor</td>
<td>18</td>
<td>27%</td>
<td>16</td>
</tr>
<tr>
<td>XP_007607154.1</td>
<td>E3-Ubiquitin Ligase</td>
<td>15</td>
<td>158%</td>
<td>23</td>
</tr>
<tr>
<td>XP_00760958.1</td>
<td>Sister chromatid cohesion protein PDSS</td>
<td>11</td>
<td>11%</td>
<td>35</td>
</tr>
<tr>
<td>XP_007610386.1</td>
<td>Alpha-galactosidase A</td>
<td>10</td>
<td>168%</td>
<td>8</td>
</tr>
<tr>
<td>XP_007614637.1</td>
<td>Heat shock-related 70 kDa protein 2</td>
<td>9</td>
<td>10%</td>
<td>11</td>
</tr>
<tr>
<td>XP_007633232.1</td>
<td>Lysosomal acid lipase</td>
<td>7</td>
<td>35%</td>
<td>8</td>
</tr>
<tr>
<td>XP_007608107.1</td>
<td>Peptidyl-prolyl cis-trans isomerase B</td>
<td>4</td>
<td>17%</td>
<td>11</td>
</tr>
<tr>
<td>XP_007634117.1</td>
<td>Protein disulfide-isomerase</td>
<td>4</td>
<td>85%</td>
<td>4</td>
</tr>
<tr>
<td>XP_007646887.1</td>
<td>Delta-1-pyrroline-5 synthetase</td>
<td>3</td>
<td>56%</td>
<td>4</td>
</tr>
<tr>
<td>XP_007622992.1</td>
<td>Ubiquitin</td>
<td>3</td>
<td>22%</td>
<td>4</td>
</tr>
<tr>
<td>XP_007612874.1</td>
<td>Heat shock 70 kDa protein 1A</td>
<td>3</td>
<td>31%</td>
<td>1</td>
</tr>
<tr>
<td>XP_007610885.1</td>
<td>Uncharacterized protein LOC100755734</td>
<td>3</td>
<td>35%</td>
<td>4</td>
</tr>
<tr>
<td>NP_001233694.1</td>
<td>Peroxiredoxin-1</td>
<td>2</td>
<td>37%</td>
<td>1</td>
</tr>
</tbody>
</table>

†While a peptide unique to this protein was identified with high confidence, it is not clear whether the entire protein was present in the sample or a clipped portion.
Quantitation was used to quantitate low level HCPs for which only one peptide was detected. Table 3.2 presents the average intensity measurements and standard deviation for the 5 protein standards. The proteins standards at the 0.5 ppm and 2.5 ppm levels were identified by one peptide and the remaining protein standards were detected by several peptides (but quantitated using the highest intensity peptide). The exclusion of myoglobin produced a linear ($R^2 > 0.990$) calibration curve that was used for assessment of the HCP concentration. Myoglobin was detected by large, low responding peptides, and its concentration was underestimated by a factor of 2 based on the linear calibration plot.

As noted above, Experiment 1 used 10 minute retention time windows in PRM to ensure detection of putative peptides for verification. Three technical replicates were used to obtain an initial estimate of the concentration levels, along with repeatability. Experiment 2, which quantitated HCPs in a second digestion of the purified mAb was conducted to assess the reproducibility, accuracy, and linearity of the label-free PRM quantitative strategy. Here, 6 minute retention time windows were used to increase the number of data points per analyte peak, resulting in generally more reproducible measurements (Table 3.1). The agreement in HCP concentration between Experiments 1 and 2 provides an indication of reproducibility.

The accuracy of quantitation was examined in Experiment 2 using ten SIL peptides that spanned the retention range of the gradient. The peptides were spiked into five aliquots of the digested purified mAb at 0, 2.5, 5, 10, or 20 fmol/injection and analyzed alongside the HCP peptides by PRM in duplicate. The spiked peptide concentrations were plotted against the concentrations measured using the calibration curve (generated from the protein standards). Nine peptides showed linear ($R^2 > 0.990$) response in the 2.5 – 20 fmol range. The linear regression slopes for each peptide ranged from 0.4 to 2.4, indicating that this quantitative strategy should be accurate to within a 2.5 fold difference. This range of error is likely the result of differences in peptide response factors and/or ion suppression from co-eluting mAb peptides. Quantitation based on SIL homologues is more accurate because both the SIL and the endogenous peptides are subjected to the same ion suppression effects. However, in a practical setting, these standards may not be readily available, and this rough label-free approximation may be sufficient in many applications. LC-MS measures protein concentration, but conventionally the HCP content is reported as
ppm relative to the mAb. If we assume an average molecular weight of 50 kDa, the total HCP content of approximately 500 ppm is detected in Experiment 2. Two HCPs at approximately 90 ppm correspond to 40% of the total content. The other HCPs are under 40 ppm, with 8 below 10 ppm. The generated information can be employed by downstream process engineers to monitor specific HCP impurities during optimization of the purification procedures. Most importantly the transition from comprehensive HCP profiling to quantitative assessment and specific identification requires only hours.

<table>
<thead>
<tr>
<th>Protein standard</th>
<th>M.W.</th>
<th>Top peptide peak area (million intensity units)</th>
<th>Spiked Concentration</th>
<th>Measured Concentration</th>
<th>Verified peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lysozyme (chicken)</td>
<td>16 kDa</td>
<td>41 ± 2.2</td>
<td>100</td>
<td>92</td>
<td>92.0</td>
</tr>
<tr>
<td>Myoglobin (horse)</td>
<td>17 kDa</td>
<td>9.5 ± 1.9</td>
<td>50</td>
<td>44</td>
<td>21.3</td>
</tr>
<tr>
<td>Lactoferrin (bovine)</td>
<td>78 kDa</td>
<td>2.7 ± 0.2</td>
<td>25</td>
<td>4.8</td>
<td>6.1</td>
</tr>
<tr>
<td>Cytochrome C (horse)</td>
<td>12 kDa</td>
<td>0.3 ± 0.2</td>
<td>2.5</td>
<td>3.2</td>
<td>0.7</td>
</tr>
<tr>
<td>β-casein (bovine)</td>
<td>25 kDa</td>
<td>0.005 ± 0.003</td>
<td>0.5</td>
<td>0.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 3.2 – Testing of PRM quantiation with protein standards

3.5 Conclusion

This paper has presented a comprehensive DIA to PRM workflow for rapid HCP identification and quantitation in biopharmaceutical products. While DIA analysis has previously been implemented in HCP characterization using the MS²⁰ and SWATH⁴⁷ strategies, our workflow exploits the high resolution and mass accuracy of the Orbitrap mass analyzer and a novel data analysis strategy to identify HCPs at low ppm levels in the background of the mAb using 1D-LC. While HCP abundance is dependent on the downstream purification process and the specific therapeutic protein, previous studies identified a similar number of HCP species in processed mAbs.¹³ Using spiked-in protein standards as internal calibrants, the PRM verification stage allowed quantitative assessment of HCP concentration. Overall the workflow demonstrated potential as a complement to ELISA in biopharmaceutical production, especially during downstream process development. The combined targeted and untargeted workflow with PRM
verification and quantitation is not limited to HCP analysis and can be implemented for more complex proteomic samples by identifying most species by DIA and using PRM to verify potentially ambiguous identifications.

3.6 Detailed description of targeted assay library assembly and DIA data processing

This section describes two scripts that were developed for data processing in the HCP analysis workflow by connecting open source packages. The **Targeted Assay Library Assembler** performs a protein sequence database search on multiple DDA data files and generates a retention time (RT) normalized MS assay library appended with decoy assays in the OpenSWATH TraML format (Figure 3.4). The **DIA Data Analysis** script performs a combined targeted and untargeted search of the DIA data and generates a list of putative peptides formatted as a QExactive inclusion list (Figure 3.5). The scripts were written in PERL (version 5.18 or higher) and combined the functionalities of msconvert from the ProteoWizard suite, SearchGUI, DIA-Umpire, OpenSWATH from the OpenMS suite, and PyProphet. The scripts use an in-house developed mzIdentML parser, which converts the output of SearchGUI’s PeptideShaker to an assay library in an OpenSWATH compatible (tab separated values) TSV format. The mzIdentML parser was initially necessary because PeptideShaker did not export results in the pepXML format. Current versions of PeptideShaker have this function, and hence SpectraST can be used to generate the assay libraries. All of the operations can be completed using the listed software tools individually, but the scripts connect them together and thus accelerate data processing.
3.6.1 Targeted assay library assembler

3.6.1.1 Required software

- Msconvert (ProteoWizard version 3.0.9 or higher):
  http://proteowizard.sourceforge.net/tools/msconvert.html
- OpenSWATH (OpenMS version 2.1 or higher): https://github.com/OpenMS/OpenMS
- SearchGUI (version 2.9.0 or higher, requires Java version 1.7 or higher) and PeptideShaker (version 1.7.6 or higher): https://github.com/compomics/searchgui
- pyProphet (version 0.18.1 or higher, requires Python version 2.7.6 or higher):
  https://pypi.python.org/pypi/pyprophet

3.6.1.2 Script input

- Raw DDA data files from analysis of samples containing the specified retention time (RT) standards.
- SearchGUI parameter file.
- Protein sequence database in FASTA format.
- Targeted Assay Library Assembler parameter file (Figure 3.6).

3.6.1.3 Script output

Primary output:

- Database search results from PeptideShaker in mzIdentML format.
- OpenSWATH compatible retention time normalized targeted assay library appended with decoy assays.
- OpenSWATH compatible retention time standards assay library for targeted analysis alignment.

Secondary Output:

- Result files from each search engine for each data file.
• List of peptides identified at specified FDR threshold.

Targeted assay library generation script

Raw DDA data
Sample containing LC retention time standards

Msconvert (ProteoWizard)
Centroided 64-bit mgf

SearchGUI
Myrimatch & MSGF+
PeptideShaker

mzidML peptide ID file

In-house mzidML parser
1. Filter PSMs to 1% FDR.
2. Average calibrant RTs.
3. Align all runs based on average calibrant RT.
4. Select best PSM per peptide charge state.
5. Extract 10 highest intensity b and y ions.

OpenSWATH (OpenMS)
ConvertTSVtoTraML
AssayGenerator
DecoyGenerator

Targeted RT normalized assay library with Decoys

mAb peptide RT assay library

Combined, non-redundant, assay library (TSV)

Figure 3.4 – Flow diagram for automatic assembly of targeted assay library.
3.6.1.4 Script procedure

1. Each data file is converted into centroided 64-bit Mascot Generic Format (MGF) using msconvert.

2. A database search is performed on each MGF file by SearchGUI using the parameters indicated in the SearchGUI parameter file.

3. The database search results from all data files are combined in PeptideShaker, and a list of peptides is generated where the false positive IDs are limited to the specified FDR (e.g. 1%) and not compounded from combination of search results from multiple runs.

4. The database search results are processed in PeptideShaker individually for each MGF file, and exported in the mzidentML format.

5. All mzidentML files are processed using the mzidentML parser as follows:
   a. The measured retention times of the specified RT standards are averaged across all runs.
   b. A linear regression curve is generated for each result file to normalize the retention time across all runs to the averages calculated in 5a.
   c. The 10 highest intensity b or y-ion transitions are extracted for each peptide (in each charge state) identified above the set PeptideShaker confidence score (e.g. 100% confidence for targeted library) and copied into an OpenSWATH compatible TSV file.
   d. The retention times for each assay are normalized based on the calibration curves generated in step 5b.
   e. Redundant identifications are removed, retaining only the matches with the highest PeptideShaker raw scores for each identified peptide charge state.

6. The assay library is filtered against the peptide list from Step 3, and only peptides identified at the true FDR threshold are retained.

7. The assays for RT standards are copied into a separate TSV file to be used for retention time calibration during DIA data analysis. At this step, a different set of peptides can be selected, in our case peptides from the mAb were used for RT normalization.
The targeted assay library and the RT standards library are converted from the TSV format into the OpenSWATH TraML format using the ConvertTSVtoTraML and the AssayGenerator scripts from the OpenSWATH toolset.

The assay library is appended with decoy assays using the OpenSWATH DecoyGenerator script and is ready to be used in targeted DIA data analysis.

3.6.2 DIA data processing

3.6.2.1 Required software

- Msconvert (ProteoWizard version 3.0.9 or higher):
- OpenSWATH (OpenMS version 2.1 or higher): [https://github.com/OpenMS/OpenMS](https://github.com/OpenMS/OpenMS)
- SearchGUI (version 2.9.0 or higher, requires Java version 1.7 or higher) and PeptideShaker (version 1.7.6 or higher): [https://github.com/compomics/searchgui](https://github.com/compomics/searchgui)
- pyProphet (version 0.18.1 or higher, requires Python version 2.7.6 or higher): [https://pypi.python.org/pypi/pyprophet](https://pypi.python.org/pypi/pyprophet)
- DIA Umpire (version 2.0 or higher): [http://diaumpire.sourceforge.net/](http://diaumpire.sourceforge.net/)

3.6.2.2 Script input

- Raw DIA data files.
- Targeted assay library (generated by the Targeted Assay Library Assembler).
- RT standards assay library (generated by the Targeted Assay Library Assembler).
- SearchGUI parameter file (use same file as in Targeted Assay Library Assembler).
- Protein sequence database in FASTA format (use same file as in Targeted Assay Library Assembler).
- DIA Data Analysis script parameter file (Figure 3.7).
- DIA Umpire parameter file.
3.6.2.3 Script output

Primary Output:

- QExactive compatible inclusion list containing HCP peptides.
- DIA data converted into pseudo-MS2 files (3 MGF files for each DIA data file).
- OpenSWATH Workflow result files for the targeted and untargeted libraries.
- SearchGUI database results of the pseudo-DDA data.
- Untargeted assay library.

Secondary Output:

- Result files from individual search engines.
Figure 3.5 - Flow diagram for automated DIA data processing.
3.6.2.4 Script procedure

1. Each DIA data file is converted into 64-bit mzXML format (the MS1 data is kept in profile mode) by msconvert.

2. The DIA files in mzXML format are processed with the OpenSWATH Workflow using the targeted assay library and the RT standards library.

3. The DIA files are then converted into pseudo-DDA files (MGF format) by DIA Umpire using the parameters indicated in the DIA Umpire parameter file.

4. Each generated MGF file is searched in SearchGUI and PeptideShaker individually and the results are exported as separate mzidentML files.

5. The files are processed using the mzidentML parser, as in the Targeted Assay Library Assembler, except that a lower PeptideShaker confidence threshold and hence a higher FDR (e.g., 5%) is used to filter identifications. The retention times are calibrated using the RT standards library.

6. The mzidentML parser generated “untargeted” assay library is converted using the ConvertTSVtoTraML, AssayGenerator, and DecoyGenerator scripts as in the Targeted Assay Library Generator.

7. The DIA files in mzXML format are processed with the OpenSWATH Workflow using the “untargeted assay library” and the RT standards library.

8. The results from the untargeted and targeted OpenSWATH searches are combined and scored using pyProphet, which identifies the best peak for each assay.

9. All matched peaks from pyProphet are exported to produce a non-redundant list of putative peptides in all detected charge states.

10. The list is filtered to exclude a set of proteins specified in the script parameter file (e.g., trypsin and the therapeutic protein).

11. The putative peptide list is manually pasted into a PRM method file for PRM verification and quantitation.
Figure 3.6 – Targeted assay library Generator parameter file.

The listed parameters match those used in the investigation.
Figure 3.7 – DIA data analysis parameter file.

The listed parameters match those in the investigation.
3.7 References


Chapter 4: Combination of DDA and DIA in a Preliminary Investigation of Extracellular Vesicles Extracted from Blood Plasma by Size Exclusion Chromatography

This chapter contains excerpts from "Mass Spectrometry-Based Molecular Characterization of Extracellular Vesicles: Lipidomics and Proteomics," which was primarily written by the dissertation author with contributions from Arseniy Belov, Dr. Ionita Ghiran, Dr. David Frank, and Dr. Shashi Murphy. The presented proteomics study was conducted by the dissertation author, the samples were gathered and donated by Natasha Barteneva’s group and the nano-flow cytometry analysis was carried out by Dr. John Tigges. The preliminary study and publication of the review were supervised by Dr. Alexander Ivanov.
4.1 Abstract

Extracellular vesicles (EVs) is an umbrella term for anuclear, phospholipid bilayer enclosed particles including multi-vesicular body derived exosomes (30 - 100 nm diameter), membrane shed microparticles (100 - 1000 nm diameter), apoptotic blebs, and other vesicular particles. EVs contain proteins, lipids, and RNA and have been observed to be ubiquitously shed from all cell types. Although once considered a mechanism for disposal of cellular debris, the role of EVs in intercellular communication has been illuminated over the last decade. EVs have been shown to regulate various physiological phenomena including inflammation, immune response, and cancer progression. Isolation of EVs from physiological fluids is an opportunity to intercept the chemical messages contained therein and to discover novel biomarkers that can be monitored without invasive surgery. EVs collected from blood are of special interest because they are potentially representative of EVs shed from all tissues in the body. However, LC-MS based molecular profiling of EVs isolated from blood plasma is hampered by potential contamination from serum proteins and non-EV particles such as lipoproteins and protein aggregates.

This chapter introduces EVs, discusses potential sources of contamination and conventional isolation strategies. An EV isolation workflow based on size exclusion chromatography (SEC) is presented as a rapid approach that does not require expensive instrumentation. The major challenge of EV isolation by SEC is the inherent variance of serum protein concentration within and between individuals, which results in EV samples of variable purity. The high intensity bias of DDA leads to stochastic identification of low abundance species, which results in limited coverage of samples containing high levels of contaminants. The reproducibility of DIA is leveraged to complement DDA identifications across a small set of patient and control samples. This preliminary investigation (1) confirmed that EVs are enriched using the SEC isolation procedure, (2) suggested that supplemental DIA analysis can normalize some of the variance of DDA analysis, and (3) identified several differentiated proteins between colorectal cancer patients and controls. The section is concluded with a look forward at further development of rapid EV isolation and analysis strategies that can be feasibly implemented in clinical settings.
4.2 Introduction

This section is composed of updated excerpts from a paper I co-authored “Mass Spectrometry-Based Molecular Characterization of Extracellular Vesicles: Lipidomics and Proteomics.”

The ability to detect disease early in its progression is necessary to improve odds for successful treatment, survival, and quality of life. It is increasingly clear that there is an urgent need to develop new, non-invasive or minimally invasive diagnostic assays for early disease detection, disease risk assessment, and patient therapy response monitoring. Extracellular vesicles (EVs) have been isolated from blood plasma and serum, urine, pleural effusions, bronchial fluid, synovial fluid, ocular fluids, breast milk, human saliva, and cerebrospinal fluid. The International Society of Extracellular Vesicles (ISEV), the American Society for Exosomes and Microvesicles (ASEMV), and the Journal of Extracellular Vesicles have been established to advance the field of EV research. Furthermore, three curated data repositories have been established to help disseminate proteomic, lipidomic, and transcriptomic EV data: ExoCarta, Vesiclepedia, and EVpedia. Much of the interest in EVs stems from their accessibility from physiological fluids and their proven involvement in biological processes such as tumor niche formation, metastasis, and immune system modulation, which highlights their potential as minimally invasive biomarker sources.

4.2.1 Nomenclature and EV categories

The term EV was coined to encompass extracellular phospholipid bilayer membrane-enclosed vesicles, which were previously described as vaults, microsomes, lipid vesicles, ectosomes, microvesicles, microparticles, exosomes, and oncosomes. In many publications, distinctions are made between multi-vesicular body-derived exosomes with a diameter of 30 - 100 nm and membrane-shed microparticles with a diameter greater than 100 nm (sometimes the upper size limit of 1000 nm diameter is suggested). However, due to an overlap in physical properties (see Table 4.1 and Figure 4.1), isolation via various preparative techniques may result in the extraction of both sets of particles. The EV term has also been used to describe even larger vesicular structures called large oncosomes (1,000 - 10,000 nm) and other submicron vesicular structures. Some EV populations have been
described using function or origin related nomenclature. For example, prostasomes are most broadly defined as vesicles presenting in seminal fluid. Although “true” prostasomes are particles secreted exclusively by the prostate, EVs isolated from seminal plasma also include vesicles secreted at other stages of the reproductive tract. As of yet, systematic molecular profiling of homogeneous EV subpopulations of a specific type has not been accomplished. Thus the EV research community encourages the use of the umbrella term EV to describe the variety of vesicular structures in corresponding isolates.

4.2.2 Isolation of EVs and potential sources of contamination

The ISEV has addressed the heterogeneity of EV isolates in a recent position statement, where a minimal set of experimental requirements for distinguishing EV from non-EV samples was presented. In summary, the enrichment of EVs must be confirmed by tracking known EV-associated protein markers relative to the source fluid and by visualization with transmission electron microscopy (TEM) or atomic force microscopy (AFM). Heterogeneity of EV size distribution can be measured by nano-tracking, light scattering, and resistive pulse sensing. However, these techniques do not distinguish EVs from non-vesicular particulates and should thus be used in conjunction with TEM or AFM. After the EVs presence is proven, their function can be studied with the use of appropriate negative controls. Finally, to attribute an EV function to any specific molecular species, the association of the molecular species with that EV has to be verified. These are excellent criteria for the study of EV biology; however, in a typical LC-MS molecular profiling experiment, hundreds to thousands of molecular species can be identified, and EV association is confirmed only for a small subset of species (if any) by orthogonal means. In the context of broad proteomic profiling, the preparative strategy with the related biases and contaminants, in addition to the sensitivity of the applied analytical platform, will define the EV population.

The optimal EV isolation strategy depends on the source fluid and the ultimate experimental goal. EV proteomic profiling is challenging in complex and high dynamic range biological fluids, e.g. blood plasma, due to potentially high levels of contamination by, exogeneous, non-EV proteins and other particulates. EV preparation is simpler in cell-conditioned media, where the concentration difference between EV and non-EV biomolecular species is less dramatic. For example, a recent study compared EV isolation by
ultracentrifugation, Opti-Prep density gradient flotation, and EpCAM affinity magnetic bead based pull-down from LIM1863 colorectal cancer cell line conditioned media. The study concluded that affinity purification most efficiently isolated EVs from the media proteins. Affinity purification, however, is not ideal for EV isolation from complex high dynamic range biological sources such as blood plasma for downstream proteomic profiling. Even low cross-reactivity of the pull-down antibody or non-specific binding of the beads can overwhelm the sample with plasma proteins. However, affinity purification of EVs from blood plasma has been effective in discovery of micro RNA marker candidates for ovarian cancer, because protein contamination was not detrimental.

As opposed to the above mentioned results of EV isolation from cell culture media, a recently conducted comparison of EV isolation strategies from blood plasma (ultracentrifugation, OptiPrep density gradient, and EpCAM affinity pull-down) found that the density gradient produced the cleanest samples. Another comparison of six protocols for isolating EVs from urine found that a modified precipitation protocol yielded the most EVs and EV mRNA, but the ultracentrifugation-based protocols were better for downstream proteomic analysis. A more recent comparison of preparative techniques suggested that size-exclusion chromatography (SEC) produced EV samples of similar purity to density gradient isolation, and this strategy is investigated further in the present chapter. The most popular preparative technique is currently differential centrifugation followed by high gravitational force (on the scale of 100,000 x g) ultracentrifugation to sediment EVs. Additional selection by a sucrose gradient is often used to purify particles in the 1.13-1.21 g/mL density range, and these isolations have been defined as the exosome fraction (although it is uncertain if such preparations contain exclusively endocytic, multi-vesicular body (MVB) derived vesicles that are biologically defined as exosomes). Ultrafiltration may be used in conjunction with or instead of centrifugation to eliminate various contaminant particles based on differences in their hydrodynamic radii. It is important to note that the throughput of ultracentrifugation based EV purification is limited by the rotor space (6 - 8 samples/rotor) and requires several hours for each set of samples. Table 4.1 summarizes the various particles that can potentially contaminate EV isolations due to overlap in physical properties, as illustrated in Figure 4.1. Furthermore, isolation strategies differ between laboratories with little consensus on the purity of the isolated vesicle-containing samples, and little considered factors such as ultracentrifuge rotor design can influence EV yield.
As shown in Table 4.1 and Figure 4.1, some virions such as HIV-1\textsuperscript{41, 42} fall within the size range of exosomes and can thus contaminate EV samples.\textsuperscript{45} Low density lipoprotein particles (LPs) have a similar diameter to exosomes, and high density lipoproteins (HDL) overlap in density with all classes of EVs.\textsuperscript{43} Thus, it is highly likely that some isolation protocols will also purify lipoproteins. A 2010 survey of ExoCarta showed that apolipoprotein E (ApoE) was identified in 5 of 19 proteomic studies.\textsuperscript{35} Further survey of Vesiclepedia also lists apolipoproteins A1, A2, A4, B, C1, C3, D, H, M, and O.\textsuperscript{12} A review of proteomic studies of HDL particles has shown that while some proteins attributed to EVs like gelsolin are also found in HDL preparations, EV characteristic proteins such as 14-3-3, tetraspanins, and integrins were not observed.\textsuperscript{46} Further overlap between lipoproteins and EVs has been shown, where tumor necrosis factor receptor 1 (TNFR1), a marker of exosomes,\textsuperscript{47} has been identified in a population of blood circulating vesicles 27-36 nm in diameter that sediment at 1.1 g/mL, but do not contain the exosome signature lipid raft macrodomains.\textsuperscript{48} These particles have been termed as exosome-like vesicles and were shown to co-migrate with low density LPs during gel filtration fast protein liquid chromatography (FPLC) - a chromatography technique that separates analytes by size. However, rate zonal centrifugation was used to partially separate the two populations of particles (LDLP sedimented at peak density of 1.03 g/mL, and TNFR1 bearing exosome-like vesicles sedimented at peak density at 1.09-1.11 g/mL. Furthermore, the study found that LAMP-1, LAMP-2, ICAM-1 and TfR (also common exosome markers) co-segregated with high density LP, which further shows that lipoprotein particles can contribute to proteomic and lipidomic EV profiles in this and other studies.
Figure 4.1 – Overlapping physical properties of biological vesicles and microparticles.\(^1\)
Overlaps in the physical properties of plasma vesicles complicate the isolation of homogeneous EVs, while structural properties and characteristic molecular markers help define vesicle categories.

Predominant contamination by non-vesicular proteins from the source fluid is a further challenge limiting the depth EV proteomic analysis. Contaminating protein derived peptides suppress ionization and hinder LC-MS detection of low abundance peptides produced from EV proteins. In urinary EV isolation, THP (Tamm-Harsfall protein) may polymerize into a disulfide-linked fibrous network that entraps EV particles and precipitates the particles at unusually low centrifugation velocity (17,000 x g instead of >100,000 x g).\(^4\)\(^9\) Non-polymerized THP is also a major component of the high gravity EV pellet. The addition of DTT during the isolation step has been shown to reduce loss of EV sample in the low gravity centrifugation pellet by breaking apart the THP networks, but the approach was ineffective in removing the protein from the high gravity pellet.\(^4\)\(^9\) Additionally, the presence of reducing agents in the isolate can lead to protein denaturation and increased aggregation. Insoluble immunocomplexes can form by
aggregation of IgG, IgM, and associated proteins during preparation of plasma derived EV samples. These aggregates are prevalent in patients with rheumatoid arthritis and other autoimmune disorders. Immunocomplexes are also a potential contaminant in healthy controls due to overlaps in size range. Such complexes introduce high levels of immunoglobulins to the high gravity EV pellets,\textsuperscript{40} thus reducing detection of EV proteins. Even without aggregation, high abundance proteins infiltrate EV samples in minimal volumes of left-over supernatant or by adhering to container walls or EVs themselves. Protein contamination can be mitigated by additional sample purification steps such as density gradient floation, or by immunoaffinity isolation of EV subpopulations.\textsuperscript{50, 51} Separation of intact EVs by hydrodynamic radius can help isolate subpopulations among these particles, while reducing the extent of contamination. For example, field flow fractionation was used to separate EVs into five fractions by size.\textsuperscript{52} Size exclusion chromatography (SEC) has also been used to separate EVs from argonaute 2 complexes (stable extracellular mRNA and protein complexes that can interfere with transcriptomic profiling).\textsuperscript{38} More recently, and as discussed in this chapter, SEC has demonstrated utility as a stand-alone technique for isolation of EVs from source fluid proteins.\textsuperscript{53}
<table>
<thead>
<tr>
<th>Particle Name</th>
<th>Description</th>
<th>Diameter (nm)</th>
<th>Density (g/mL)</th>
<th>Markers</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exosomes</td>
<td>Vesicles released from MVBs</td>
<td>30-120</td>
<td>1.13-1.21</td>
<td>Alix, TSG101, HSC70, CD63, CD81, CD9, amongst others</td>
<td>Mathivan (2010)^35</td>
</tr>
<tr>
<td>Ectosomes, Microparticles</td>
<td>Vesicles formed by direct budding from cell membrane</td>
<td>50-1000</td>
<td>~1.16</td>
<td>Surface phosphatidylserine, Selectins, integrins, CD40, metalloproteases</td>
<td>Barteneva (2013)^35</td>
</tr>
<tr>
<td>Apoptotic Blebs</td>
<td>Apoptosis fragments</td>
<td>50-500</td>
<td>1.16-1.28</td>
<td>Surface phosphatidylserine, Histones, Calnexin, Cytochrome C</td>
<td>Mathivan (2010)^35</td>
</tr>
<tr>
<td>Large Oncosomes</td>
<td>Non-apoptotic plasma membrane blebs shed by “amoeboid” migrating tumor cells.</td>
<td>1000-10,000</td>
<td>N/A</td>
<td>Cav-1, ARF6</td>
<td>Di Vizio (2012)^26</td>
</tr>
<tr>
<td>Argonaute Complexes*</td>
<td>Secreted RNA-Protein complexes</td>
<td>~12</td>
<td>N/A</td>
<td>Ago 2, miRNA</td>
<td>Arroyo (2011)^38</td>
</tr>
<tr>
<td>Immuno-complexes</td>
<td>Protein clusters containing immunoglobulins</td>
<td>50-250</td>
<td>&gt;1.21</td>
<td>Immunoglobulins</td>
<td>Gyorgy (2011)^40</td>
</tr>
<tr>
<td>Viral Particles</td>
<td>Influenza</td>
<td>85-120</td>
<td>N/A</td>
<td>Hemagglutinin</td>
<td>Roy (2000)^61</td>
</tr>
<tr>
<td></td>
<td>HIV-1</td>
<td>113-139</td>
<td>1.16-1.18</td>
<td>Gp120, gp41, gp32, Host MHC-II</td>
<td>Zhu (2003)^52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cantin (2007)^54</td>
</tr>
<tr>
<td>Lipoprotein Particles (LP)</td>
<td>Very Low Density LP-1</td>
<td>33-70</td>
<td>&lt;1.006</td>
<td>Apolipoprotein B-100</td>
<td>Berneis (2002)^43</td>
</tr>
<tr>
<td></td>
<td>Very Low Density LP-2</td>
<td>30-33</td>
<td>1.006-1.010</td>
<td>Apolipoprotein E</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermediate Density LP-1</td>
<td>28.5-30</td>
<td>1.008-1.022</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermediate Density LP-2</td>
<td>27.2-28.5</td>
<td>1.013-1.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low Density LP</td>
<td>22-27.2</td>
<td>1.019-1.060</td>
<td>Apolipoprotein B-100</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1 – Potential Non-EV contaminants.
Particles of various types and physical attributes may be co-isolated in MP or exosome preparations. As there are slight discrepancies between the physiological properties of particles between publications, this table presents the broadest ranges found. *Argonaute complexes are known contaminants in RNA research of EVs. The physical properties provided here are derived from analysis of Ago2 in complex with miR-20a X-Ray structure (PDB entry 4F3T).^38,39
4.2.3 Proteomic profiling of EVs.

EVs derived from different sources tend to retain unique signatures and markers of their origin, a feature that can be utilized for diagnostic and therapeutic purposes. Proteomic profiling of EVs has a dual purpose: (1) exploitation of EVs for relatively non-invasive collection and enrichment of disease-associated proteins to be used as prognostic and diagnostic biological markers and (2) investigation of the roles of EVs and their proteomes in biological phenomena. Aside from challenges in isolating pure EV populations, there are several sample-specific challenges in proteomic analysis of EVs in physiological fluids. As previously discussed, high abundance proteins from the source media contaminate EV isolates. EV recovery and hence protein yield can be low and is further reduced due to losses at various stages of isolation, sample processing, and fractionation of proteomic constituents. Furthermore, the expected wide dynamic range of cell-type specific EV concentrations and the overrepresentation of EV populations derived from several predominant cell types (such as red blood cells and platelet EVs in blood) can result in suppression of signal from rare EVs and related proteins from less abundant cell types. Biological responses to extraction of EVs from whole blood, such as platelet activation, inflammation, and coagulation, can introduce further artifacts to EV samples. Similar considerations are required for isolation of EVs from breast milk. Additionally, membrane-bound and embedded proteins, which are notoriously difficult to recover in bottom-up proteomic analysis, comprise an estimated 28% of biological fluid derived EVs and 34% of cell media-derived EVs.

Due to all these challenges, there is a high variability in the number of protein identifications between studies that can be attributed to either differences in sensitivity of analysis, differences in isolation protocols, and/or to differences in complexity and dynamic range of samples. For instance, it would not be surprising to find aberrant incorporation of proteins into cancer cell line EVs, where other processes are also abnormal, and it would also not be surprising to identify a relatively modest repertoire of proteins in plasma isolations dominated by platelet shed EVs. Additionally, the composition of biological fluids can vary drastically throughout the day due to circadian rhythms and environmental factors, introducing further variability in EVs and potential contaminating species. Given these complications, time-consuming, sophisticated isolation strategies may not be scalable to clinical investigations. The SEC
based isolation process investigated in subsequent sections presents a rapid and simple alternative, which with further development could provide a robust high-throughput EV analysis.

4.2.4 SEC based EV isolation

Conventional EV isolation techniques by ultracentrifugation and density gradient centrifugation\(^{30}\) are low-throughput and impractical in a clinical setting. Alternatively, SEC, introduced in Chapter 1 can be used to isolate EVs from serum proteins in minutes without specialized equipment.\(^{53}\) SEC uses porous low binding beads to allow large particles (i.e. EVs), which cannot enter the pores, to readily pass through the bed of the stationary phase, while smaller sample constituents (e.g. proteins) are slowed by the tortuous path through the pores. SEC is a low-resolution chromatographic technique, and SEC-based EV isolations may thus contain other particles of similar size (30 - 1000 nm) and low levels of plasma protein complexes and aggregates.\(^{65}\) While higher purity isolation can be obtained using more time and resource consuming procedures, the presented approach will be shown to be capable of rapid enrichment of EVs from blood plasma and may be applied to other highly complex physiological fluids, and samples of other types (e.g. cell culture media). The approach is diagrammed in Figure 4.2.
Figure 4.2 – Workflow implementing rapid SEC based isolation of EVs.
Rapid SEC based EV isolation, verification of EV enrichment by nano-cell flow cytometry or alternative means, and parallel proteomic and optional lipidomic analysis.

4.3 Materials and Methods

4.3.1 Materials

Blood was drawn by Dr. Natasha Barteneva’s group using the Vacutainer® UltraTouch™ Push Button Blood Collection Set consisting of a 21 gauge 0.75” needle, 7” tubing (pt# 367393) and 10 mL EDTA-coated Vacutainer® tube (pt# 367856) (BD Life Sciences, Franklin Lakes, NJ). HPLC-grade solvents (water, isopropanol) and sterile Dubellco phosphate buffered saline (DPBS), pH 7.4 (Sigma-Aldrich, St. Louis MO) were filtered through a 0.22 µm filter to mitigate possible contamination with the microparticulate material. A slurry of cross-linked Sepharose CL - 2B particles (Sigma-Aldrich pt.# CL2B300) shipped in 20% ethanol were packed in rinsed 10 mL columns with a 13 mL reservoir from Pierce (pt.# 89898) to make the single use gravity-driven SEC columns. Before injection onto the column, the cell-free plasma was collected in a rinsed 10 mL disposable syringe with a Luer-lock adapter (e.g. BD
and filtered through a single-use DPBS conditioned 30 mm diameter 0.8 µm pore cellulose acetate membrane with a Luer adapter (e.g. Cameo syringe filters Sigma-Aldrich pt.#741675 or #741864). The EV fractions were collected in 2 mL low binding Eppendorf tubes from VWR (Radnor, PA) and processed using HPLC grade chloroform, methanol, water, 2,2,2-trifluoroethanol, and highest purity iodoacetamide (IAA), Tris(2-carboxyethyl)phosphine hydrochloride (TCEP), formic acid purchased from Sigma-Aldrich, and LC-MS grade Trypsin/Lys-C mix from Promega (Madison, WI pt.#V5073). The samples were analyzed using 25 cm long 75 µm internal diameter and 365 µm outer diameter fused silica capillary with a porous Kasil frit, polymerized in-house, and packed with 3 µm Magic C18 AQ, pore size 200 Å stationary phase. The samples were electrosprayed through a distal coated 20 µm ID pulled to 10 µm opening fused capillary tip (New Objective pt# FS360-20-10-D) connected to the end of the column with a 250 µm internal diameter Teflon sleeve (Thermo Fisher Scientific, pt# 160486).

4.3.2 Methods

4.3.2.1 Preparation of single-use SEC columns

Single use columns were prepared within a couple of hours before separation by filling 10 mL column cartridges with roughly 12 mL of 2% cross-linked agarose (CL - 2B) beads in 20% ethanol slurry. The beads were allowed to settle for 10 minutes as the solvent drained from the column, and the slurry volume was adjusted to 10 mL based on column markings. The filled columns were rinsed with 50% aqueous isopropanol and conditioned with DPBS before use.

4.3.2.2 SEC isolation of EVs

Frozen plasma samples (stored at -80 °C) were gathered and graciously donated by Dr. Natasha Barteneva from Harvard Medical School and Boston Children's Hospital. Following an IRB approved protocol, roughly 8 mL of blood were collected by percutaneous cubital venipuncture in EDTA-containing Vacutainers. The blood samples were centrifuged at 2,000 x g for 5 minutes to precipitate cells by the Barteneva group. The plasma supernatant was flash frozen in liquid nitrogen and donated to the Karger group. The samples were thawed and 2 mL of plasma were aspirated into the 10 mL Luer-lock syringe.
The DPBS conditioned 0.8 μm cellulose acetate disk filter was attached to the end of the syringe, and 1 mL of plasma were filtered through and injected onto the SEC column.

The injected volume was allowed to enter the stationary phase, and the column stopper was closed. 10 mL of DPBS were added into the reservoir. The stopper was opened, and the elution volume was measured by filling 2 mL Eppendorf tubes. The 4 - 6 mL volume was collected as Fraction 1, and 6 - 6.5 mL were collected as Fraction 2 for each sample. The samples were frozen and stored at -80 °C.

4.3.2.3 Nano-flow cytometric analysis of SEC fractions

To identify the SEC fractions containing the peak of EV elution, 2 mL fractions were collected. 10 μL aliquots from each fraction were diluted 100 fold in DPBS and analyzed by nano-cell flow cytometry on a MoFlo XDP flow cytometer with a NanoView module (Beckman Coulter, Indianapolis, IN) at Beth Israel Medical Center (Boston, MA) by Dr. John Tigges. Side and forward scatter were measured over 3 million events (an event constitutes a detectable particle passing through the flow cell), and the population corresponding to the 100 - 300 nm particle range was monitored as presented in Figure 4.3.

4.3.2.4 Processing of EV containing SEC fractions

The samples were thawed at room temperature and concentrated by SpeedVac to a volume of approximately 200 μL. 400 μL of methanol and 200 μL of chloroform were added, and the samples were vortexed for one minute and sonicated in a water bath for 5 minutes to break the EV structures. 200 μL of chloroform were added followed by 200 μL of water to precipitate the protein contents into the interphase. The samples were centrifuged at 14,000 x g for 10 min at room temperature. This procedure is known as the Bligh-Dyer extraction and is traditionally used to isolate lipid species from complex matrices. The samples were refrigerated at -20 °C (this did not freeze either phase) for 4 hours to settle the precipitated proteins in the interphase.

The methanol/water (upper) phase was discarded, and the lipid-containing chloroform (lower) phase was collected in a separate Eppendorf tube. The remaining protein containing pellet was lyophilized to dryness and reconstituted in 50 μL of 30 mM ammonium bicarbonate pH 8.0 in 50% aqueous 2,2,2-
trifluoroethanol to solubilize the precipitated proteins. 6 µL of 50 mM TCEP in 125 mM ammonium bicarbonate, pH 8, were added to bring the TCEP concentration to 5 mM, and the samples were incubated at 37 °C for 30 minutes. 7 µL of freshly prepared 100 mM IAA in 125 mM ammonium bicarbonate, pH 8, were added to bring the IAA concentration to 10 mM followed by incubation at room temperature for 45 minutes in the dark. 457 µL of 25 mM ammonium bicarbonate at pH 7.5 were added along with 2 µg of Trypsin / Lys-C Mix. Digestion proceeded at 40 °C for 16 hours. Formic acid was added to 0.1% to quench digestion, and the samples were concentrated to 50 µL volume by lyophilization.

4.3.2.5 Nano-LC-MS analysis

1 µL of sample were injected directly into the analytical column at 250 nL/min in 98% Solvent A (0.1% formic acid in water) 2% Solvent B (0.1% formic acid in acetonitrile). The separation gradient was delivered at 250 nL/min as follows: desalting for 15 min at 2% B; linear ramp to 42% B over 90 min; linear ramp to 95% B over 15 min, 10 min hold at 95% B; drop to 2% B in 1 min, hold at 2%B for 14 min. The eluting peptides were electrosprayed at 1.9 - 2.3 kV and analyzed by DDA or DIA on a QExactive mass spectrometer (Thermo Fisher Scientific).

DDA analysis was carried out using the Top 15 strategy with the following settings: MS1 scans at 70,000 resolution at 200 m/z, 1,000,000 automatic gain control (AGC), maximum C-trap fill time of 110 ms; MS2 scans at 17,500 resolution at 200 m/z with AGC set to 100,000 and 50 ms maximum C-trap fill time with precursor isolation in 2 m/z wide windows and HCD at 28 normalized collision energy. The signal for dodecamethylcyclohexasiloxane (445.12 m/z) was used as an internal lock-mass calibrant. The MS1 settings were matched in DIA analysis, except the range was set to 390 – 1,100 m/z. MS2 data were acquired using 22 m/z wide isolation windows covering the 400 – 1,000 m/z range. 27 DIA scans were acquired for each MS1 scan, and adjacent isolation windows had no overlap, but there was an 11 m/z offset between acquisition cycles (scan of the entire analysis range). The DIA scans were acquired with 35,000 resolution at 200 m/z with AGC set to 100,000 and 110 ms maximum C-trap fill time.
4.3.2.6 Peptide and protein identification

All DDA data were searched together using MaxQuant (v.1.5.3.28) against the May 2016 SwissProt human protein sequence database. The MaxQuant match-between-runs feature was used to address the “missing value” issue associated with stochastic peptide and protein identifications in DDA by matching peptide peaks between runs based on high accuracy mass and retention time measurements (even if the peptide was not identified by an MS2 scan in a given run). The MaxQuant workflow started with a survey search using 20 ppm MS1 and MS2 mass tolerances with tryptic enzyme specificity and carbamidomethylation of cysteine as a static modification to recalibrate the m/z measurements off-line thus improving mass accuracy in the main search. After recalibration, the main search used 4.5 ppm precursor mass and 8 ppm fragment mass errors. Carbamidomethylation of cysteine was set as a static modification, oxidation of methionine was set as a variable modification and digestion specificity was set to semi-tryptic with two allowed missed cleavages. The results were filtered to 1% FDR at the peptide and protein levels. The DDA data were also processed in SearchGUI against the same database using the MS-GF+ search engine to allow comparison with DIA data processing. The mass errors were set to 10 ppm for precursors and fragments, and all other parameters matched the MaxQuant main search.

DIA data were analyzed first by untargeted analysis through DIA-Umpire\(^{69}\) conversion as described in Chapter 3. The converted data were searched by SearchGUI\(^{70}\) matching the parameters used in processing of DDA data. The search results from DDA analysis by MaxQuant and MS-GF+\(^{71}\) and untargeted analysis of DIA data were combined into a Skyline\(^{72}\) assay library. The Skyline library was appended with an equal number of shuffled decoy assays. All DIA data were imported into Skyline and interrogated by the cumulative assay library. mProphet\(^{73}\) peak scoring was used to identify matches using a \(q < 0.01\) cut-off. Peptides were quantitated based on the combined signal from the 3 highest intensity ions, and proteins were quantitated based on the average of 3 highest intensity peptide peak areas. Perseus\(^{74}\) was used for statistical analysis of the DDA results and combined DDA and DIA results.
4.3.2.7 Experimental design

Twenty frozen plasma samples (5 from colorectal cancer patients and 5 matched controls; 5 inflammatory bowel disease (IBD) patients and 5 matched controls) obtained from Natasha Barteneva’s group at Harvard Medical School and Boston Children’s Hospital (according to an IRB approved protocol), thawed, and processed using the described workflow. Two SEC fractions were collected for each plasma sample: Fraction 1 collected the peak of the EV elution at 4 – 6 mL elution volume, Fraction 2 was collected at the start of the plasma protein peak at 6 – 6.5 mL elution volume. Each fraction was analyzed twice using DDA and once using DIA. The fractions from each donor were processed sequentially. Donors’ samples were analyzed in random order, with a column wash performed between samples collected from different donors.

4.4 Results and Discussion

4.4.1 SEC enriches EVs and depletes serum proteins

EVs were monitored in the SEC elution volumes using nano-flow cytometry at Beth Israel Medical Center by Dr. John Tigges (Figure 4.3). The nano-flow cell allowed detection of particles below the size limit of conventional flow cytometry (500 nm). The analysis monitored forward and side scatter from a 488 nm laser and particles within the 100 – 300 nm range were counted as EVs. For each fraction, the flow cytometer acquired 3 million scatter events, the ratio of events within the side and forward scatter range associated with EV particles is presented in Figure 4.3. This experiment along with results published by other groups was used to select the optimal SEC fraction for comparative analysis. Proteomic analysis and gene ontology term enrichment analysis (Figure 4.4) confirmed that EV proteins were more prominent in the earlier fraction, indicating that low resolution separation between EVs and exogenous proteins occurred. More EV-characteristic proteins were identified in Fraction 1, and the Exosome gene ontology (GO) term was better represented (a larger fraction) in Fraction 1 compared to the adjacent Fraction 2 where the number of identified protein groups was reduced due to increased contamination from serum proteins.
Figure 4.3 – Nano flow cytometry analysis of SEC elution volumes.
Flow cytometry shows that the EV population elution peaks between 4 - 6 mL, and diminishes between 6 - 6.5 mL elution volume. The elution volume indicated in the bar graph matches the flow cytometry scatter plots above. The blue bar corresponds to the event ratio for Donor 1 and the red bar corresponds to the event ratio for Donor 2.
Figure 4.4 – GO term comparison of two adjacent SEC fractions.
More protein groups and EV associated protein groups are identified in the 4 - 6 mL SEC elution volume, because the 6 - 6.5 mL fraction has increased masking from serum proteins.

4.4.2 DDA proteomic analysis of SEC isolated EVs

All 80 DDA data files (20 plasma samples x 2 SEC fractions x 2 replicates) were analyzed together in MaxQuant with the match-between-runs feature enabled. 273 ± 38 protein groups were identified in the colorectal cancer patient samples and 299 ± 31 were identified in the corresponding controls. Significantly more protein groups were identified in the IBD cohort (470 ± 195) and corresponding controls (515 ± 180). Since all samples were processed blindly and in random order, the significant difference in protein identifications between the sets of controls is likely due to differences in plasma isolation procedures between the two sites which contributed the samples, highlighting the necessity for more uniform blood collection. 482 ± 178 protein groups were identified in Fraction 1 compared to 311 ± 199 in Fraction 2, which matched the total protein identifications trend presented in Figure 4.4. As mentioned in the introductory sections, this relatively modest depth of analysis is somewhat expected for samples dominated by platelet derived EVs, and yet some differences were observed between the cohorts.
Differentially expressed EV proteins were identified based on label-free quantitation between the patient and control samples using Perseus. First, the data were normalized across runs by expressing the intensity of each protein group as a percentage of the cumulative intensity of the run. Then, the data were log transformed (base 2) and a volcano plot (fold change vs. negative log of the p-score from a two-sided homoscedastic t-test) was generated comparing the runs in each cohort to the matched controls (Figure 4.5). The analysis found that 10 protein groups were significantly (FDR < 1%, minimum fold change (s0) set to 0.1, or 1.07 change before log 2 transformation) enriched in the colorectal cancer samples, and 9 proteins groups were reduced. These differences appeared when both SEC fractions were analyzed together and when only Fraction 1 results were considered. No protein differences were determined significant at 1%FDR in the IBD cohort compared to controls. The proteins enriched in colorectal cancer samples included innate immune response components: complement 4B, 4A, 9, and 1 subcomponent; protease inhibitors: alpha-1-antichymotrypsin, alpha-1-antitrypsin, inter-alpha-trypsin-inhibitor 4; and platelet factor 4 and serum amyloid A1. Apolipoproteins AIV, CIV, CIII, and A1, haptoglobin-related, alpha-2-macroglobulin, fibronectin, serum peroxonase, and phosphatidylinositol-glycan-specific phospholipase D were more abundant in the control samples. The p-scores and fold changes are presented in Table 4.2 along with the same values determined by DIA analysis. These preliminary data suggest that differences between cohorts can be detected with this strategy but further development is required for more robust quantitation.
Figure 4.5 – Volcano plot comparing 5 colorectal cancer patients (red) and 5 controls (blue). Based on the fold change and a two sided t-test 9 proteins are reduced and 10 are enriched in the colorectal cancer plasma samples. False discovery rate is set at 1% and s0 is the minimum fold change after the log2 transformation set to 0.1.
<table>
<thead>
<tr>
<th>Protein Accession</th>
<th>Description</th>
<th>DDA -log(p)</th>
<th>DDA Fold change</th>
<th>DIA -log(p)</th>
<th>DIA Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0DJ18</td>
<td>Serum amyloid A-1 protein</td>
<td>5.07</td>
<td>4.28</td>
<td>1.29</td>
<td>3.03</td>
</tr>
<tr>
<td>P02776</td>
<td>Platelet factor 4</td>
<td>5.59</td>
<td>2.64</td>
<td>0.12</td>
<td>0.35</td>
</tr>
<tr>
<td>P02763</td>
<td>Alpha-1-acid glycoprotein 1</td>
<td>2.95</td>
<td>1.58</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>P01011</td>
<td>Alpha-1-antichymotrypsin</td>
<td>2.90</td>
<td>1.29</td>
<td>1.42</td>
<td>0.85</td>
</tr>
<tr>
<td>B7ZKJ8</td>
<td>Inter-alpha-trypsin inhibitor heavy chain H4</td>
<td>2.55</td>
<td>1.19</td>
<td>#N/A</td>
<td>#N/A</td>
</tr>
<tr>
<td>P02748</td>
<td>Complement component C9</td>
<td>5.46</td>
<td>1.18</td>
<td>0.96</td>
<td>0.81</td>
</tr>
<tr>
<td>P01009</td>
<td>Alpha-1-antitrypsin</td>
<td>2.83</td>
<td>1.03</td>
<td>1.80</td>
<td>1.02</td>
</tr>
<tr>
<td>P0C0L4</td>
<td>Complement C4-A</td>
<td>5.29</td>
<td>0.81</td>
<td>1.38</td>
<td>0.62</td>
</tr>
<tr>
<td>P0C0L5</td>
<td>Complement C4-B</td>
<td>4.35</td>
<td>0.67</td>
<td>0.36</td>
<td>0.27</td>
</tr>
<tr>
<td>P09871</td>
<td>Complement C1s subcomponent</td>
<td>5.48</td>
<td>0.62</td>
<td>0.84</td>
<td>0.58</td>
</tr>
<tr>
<td>P01023</td>
<td>Alpha-2-macroglobulin</td>
<td>4.76</td>
<td>-0.79</td>
<td>1.69</td>
<td>-0.83</td>
</tr>
<tr>
<td>O14791</td>
<td>Apolipoprotein L1</td>
<td>8.16</td>
<td>-0.84</td>
<td>1.84</td>
<td>-0.84</td>
</tr>
<tr>
<td>P02656</td>
<td>Apolipoprotein C-III</td>
<td>3.68</td>
<td>-0.87</td>
<td>1.03</td>
<td>-0.62</td>
</tr>
<tr>
<td>P02751</td>
<td>Fibronectin</td>
<td>4.43</td>
<td>-0.87</td>
<td>0.53</td>
<td>-0.30</td>
</tr>
<tr>
<td>P27169</td>
<td>Serum paraoxonase/arylesterase 1</td>
<td>5.05</td>
<td>-0.87</td>
<td>0.82</td>
<td>-0.58</td>
</tr>
<tr>
<td>P00739</td>
<td>Haptoglobin-related protein</td>
<td>5.04</td>
<td>-0.90</td>
<td>0.76</td>
<td>0.58</td>
</tr>
<tr>
<td>P55056</td>
<td>Apolipoprotein C-IV</td>
<td>8.17</td>
<td>-0.99</td>
<td>2.36</td>
<td>-1.30</td>
</tr>
<tr>
<td>P06727</td>
<td>Apolipoprotein A-IV</td>
<td>4.00</td>
<td>-1.51</td>
<td>1.41</td>
<td>-1.49</td>
</tr>
<tr>
<td>P80108</td>
<td>Phosphatidylinositol-glycan-specific phospholipase D</td>
<td>3.57</td>
<td>-1.79</td>
<td>2.21</td>
<td>-2.03</td>
</tr>
</tbody>
</table>

Table 4.2 – Fold changes and p-scores of potentially differentiated proteins.

4.4.3 Complementary DIA analysis

DIA has been shown to identify a complementary population of proteins in previous studies. Additionally, DIA quantitation is considered more reproducible than DDA. Thus, DIA MS2 based quantitation was implemented to attempt to identify differentially expressed proteins which did not appear significant in DDA analysis due to: stochastic quantitation (beyond MS1 peak matching), lack of identification due to under-sampling, or irreproducible quantitation due to interference in the MS1 domain.

Comparison of cumulative protein and peptide identifications by untargeted DIA analysis and DDA analysis with MaxQuant and MS-GF+ (Figure 4.6) shows that more proteins/peptides were identified by DDA with a few additional species detected by DIA. This result is likely due to saturation reached after 80 analyses of very similar samples with a limited (~1,500) detectable protein repertoire. The differences between MaxQuant and MS-GF+ identifications are likely due to the more stringent MaxQuant matching criteria, fundamental differences in search engine performance, and differences in protein inference strategies. Despite adding little to the catalogue of SEC isolated EV proteins, DIA is complementary to
DDA in individual runs as demonstrated in Figure 4.7. DIA helps identify additional proteins in individual runs and allows a more thorough comparison between the two cohorts. The bar graph in Figure 4.8A confirms that reproducibility in identifications was better with DIA (more proteins were observed in all 10 donors), and Figure 4.8B shows that DIA is particularly advantageous in reproducible identification of lower intensity proteins. As shown in Table 4.2, the DIA data confirmed the direction of fold-changes for the proteins that appeared differentiated in DDA analysis, but surprisingly these changes did not appear significant based on the volcano plot, and no additional differentiated proteins were identified at the 1% FDR threshold (not shown). The statistical significance of DIA differentiation can be improved by better normalization in sample preparation and further improvements that are discussed in the following section.

![Venn Diagrams](image)

**Figure 4.6 – Cumulative protein and peptide identifications by DDA and DIA**
The Venn diagrams show the protein (left) and peptide (right) repertoires identified by DDA with MaxQuant (green), DDA with MS-GF+ (blue), and DIA using DIA-Umpire and MS-GF+ (orange).
Figure 4.7 – Filling DDA missing values with DIA
Each pair of rows (green / yellow + blue) corresponds to one donor, each column corresponds to a protein group. Green indicates that the protein was identified by an MS2 scan in at least one of two DDA runs, yellow indicates that the protein was identified by matching of MS1 peaks in MaxQuant. Blue cells correspond to DIA identifications with Skyline. For simplicity, only results from the first fraction are presented, and protein groups which were identified in fewer than 5 runs are excluded.

Figure 4.8 – Normalization of protein identifications across samples is improved with DIA
Panel A is a histogram of the number of samples (combined from both fractions) in which proteins groups were identified; DIA analysis identified proteins more reproducibly across samples. Panel B shows the number of samples in which the top 300 intensity proteins were identified sorted by intensity; DIA is particularly helpful in normalizing the identifications of low intensity (bottom 50% percentile) proteins.
4.4.4 Future directions

These preliminary results demonstrated that SEC based isolation depleted common serum contaminants such as albumin and immunoglobulins and enriched particles 100 – 300 nm in diameter. Over 1,000 proteins, including known EV markers, were identified in the isolations, but comparison of colorectal cancer patients and controls (n = 5 + 5) only found subtle differences between the two small cohorts. While the initial goal of the workflow was simplicity and quick application, improvements are required for reliable and robust quantitation, and increases in sensitivity are needed to reveal more sample differences. Future implementations of the workflow should include: normalization based on particle count, more reproducible and higher resolution SEC, and additional EV purification.

In the preliminary investigation, the same plasma volume equivalent was injected for each sample (20 µL per run), and protein abundances were normalized across runs by expressing each protein’s intensity as percentage of the summed intensity values for all detected proteins for that run. This minimal normalization was insufficient to overcome the variances of the experiment. The EV and protein concentration in plasma can vary drastically between and within individuals, thus the most logical normalization should be based on the EV concentration. EV concentration after SEC can be measured rapidly by dynamic light scattering,78,79 nano-tracking analysis of Brownian motion (e.g. Nanosight),80 size tunable pore sensors (e.g. NanoPore),81 or by nanoflow cytometry.82 Although these approaches require additional instrumentation, the size distribution and concentration of EVs can serve as additional markers. The concentration of non-EV proteins can influence detection and quantitation of EV proteins and improvements in the isolation procedure will increase depth of profiling and quantitative reliability.

Improvement of the SEC column would better separate EVs from the serum proteins. Substitution of the single-use gravity driven columns with a pressure driven column would eliminate poor reproducibility from single-use column packing, manual injection, and gravity driven separation (due to differences in sample viscosities). Furthermore, the column could be paired on-line to a detector to monitor the EV and protein fractions in real-time and collect the optimal, rather than the set, elution volume. However, care must be taken not to lyse the EVs by excessive back pressure, and the column must be thoroughly washed between uses to prevent cross contamination. Given the poor resolution of SEC, especially using
large pore particles, it is unlikely that non-EV proteins could be completely depleted by SEC alone. For even deeper profiling, an additional selection mechanism is likely necessary. Affinity purification has been shown to generate pure EV populations from conditioned cell media. While even minor cross-reactivity with serum proteins makes this strategy impractical for EV isolation directly from plasma, SEC isolation can reduce the problematic contaminants. Another potentially promising complement to SEC is protein organic solvent precipitation (PROSPR),\textsuperscript{83} where an addition of acetone to 80% causes non-EV proteins to precipitate while EV proteins, stabilized by the lipid bilayer, remain in the supernatant. This additional purification step is rapid and simple to implement, and could potentially be modified to also deplete lipids and replace the currently implemented Bligh-Dyer extraction.

4.5 Conclusion

EVs are phospholipid bilayer enclosed vesicles containing proteins, RNA, and metabolites which have been shown to influence disease progression, inflammation, immune response and other biological phenomena. EVs shed into the plasma can be extracted and analyzed without invasive surgery providing an opportunity for liquid biopsy to diagnose disease and monitor progression. However, molecular profiling, in particular proteomic analysis, of EVs from high dynamic range complex biological samples like blood plasma is complicated by the overabundance of source fluid proteins, the relatively low concentration of EVs, and contamination from non-EV particles. Given these obstacles we’ve investigated a rapid SEC-based isolation procedure, which generates somewhat crude samples, but is advantageous for its simplicity and potential compatibility with diverse clinical and biological applications. Nano-flow cytometry and proteomic analysis confirmed that EVs are enriched while serum proteins are reduced. A combination of DDA and DIA was used in a preliminary investigation comparing colorectal cancer patients and controls. Due to heterogeneity of plasma, it was expected that the depth of profiling would vary between donors, and DIA was used as a measure to normalize identifications across samples. While DIA identified additional proteins in every run and led to more reproducible detection of protein groups between donors, few significant differences were observed between the two cohorts. To improve quantitative accuracy and precision, several measures have been suggested for future iterations of the workflow.
4.6 References


82. Nolte-t Hoen, E. N. M.; van der Vlist, E. J.; Aalberts, M.; Mertens, H. C. H.; Bosch, B. J.; Bartelink, W.; Mastrobattista, E.; van Gaal, E. V. B.; Stoorvogel, W.; Arkesteijn, G. J. A.; Wauben, M. H. M.,